Text books :-

1. Gupta, S.C and Kapoor, V.K. fundamentals of mathematical statistics solution chand and sons

2. S.P. Gupta, statistical methods (Revised edition 2001).

UNIT - I

(1)

Normal Distribution

Definition :-

A random variable x is said to have a normal distribution with Parameter μ called mean and σ^2 called variance if its P.d.f is given by the Probability law,

$$f[x; \mu, \sigma] = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$$

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

(2)

Remark :-

When the random variable is normally with mean μ and standard deviation 'σ' it is customary to write x is distributed as (μ, σ^2) and is expressed by $x \sim N(\mu, \sigma^2)$

If $X \sim N(\mu, \sigma^2)$, then $z = \frac{x-\mu}{\sigma}$ is a standard normal variate with $E(z) = 0$ and $\text{Var}(z) = 1$, we write, $z \sim N(\mu, \sigma^2)$

$\therefore z \sim N(0, 1)$

The P.d.f of standard normal variate z is given by

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

$$= \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

and the corresponding distribution function denoted,

$$\phi(z) = P(z \leq z) = \int_{-\infty}^{z} \phi(u)\, du$$

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

Result - I

Prove that $\phi(-z) = 1 - \phi(z)$, $z > 0$.

Soln:-

$$\phi(-z) = P(z \leq -z)$$

$$= P(z \geq z)$$

$$= 1 - P(z \leq z)$$

$$= 1 - \phi(z)$$

Result - II

Prove that $P(a \leq x \leq b) = \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$

where, $x \sim N(\mu, \sigma^2)$

Soln:-

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

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

$$P(a \le x \le b) = P\left[\left(\frac{a-\mu}{\sigma}\right) \le \left(\frac{x-\mu}{\sigma}\right) \le \left(\frac{b-\mu}{\sigma}\right)\right] \quad (3)$$

$$= P\left[\left(\frac{a-\mu}{\sigma}\right) \le (z) \le \left(\frac{b-\mu}{\sigma}\right)\right]$$

$$= P\left[z \le \left(\frac{b-\mu}{\sigma}\right)\right] - P\left[z \ge \left(\frac{a-\mu}{\sigma}\right)\right]$$

$$= \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$$

The graph of $f(x)$ is famous bell shaped curve A top of the bell is directly above the mean μ for large value of σ, then a curve tends to flatten out and are small values of a shape Peak.

Normal distribution as a limiting form of Binomial distribution :-

Normal distribution is another limiting form of Binomial under the following conditions

i) n the number of trials is indefinitely large (i.e) $n \to \infty$ and

ii) Neither P (or) q is very small. The P.m.f of the binomial distribution with Parameters n and P is given by

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

$$= \frac{n!}{x!(n-x)!} p^x q^{n-x}; \quad x = 0, 1, 2, \cdots n \to \infty$$

Consider the standard binomial variable (1)

$$z = \frac{x - E(x)}{\sqrt{V(x)}} = \frac{x - np}{\sqrt{npq}}, \quad x = 0, 1, 2, \ldots n \rightarrow (2)$$

when $x = 0$,

$$z = \frac{-np}{\sqrt{npq}} = \frac{-\sqrt{np}\sqrt{np}}{\sqrt{npq}} = -\sqrt{\frac{np \cdot np}{np \cdot q}}$$

$$= -\sqrt{\frac{np}{q}}$$

when, $x = n$

$$z = \frac{n - np}{\sqrt{npq}} = \frac{n(1-p)}{\sqrt{npq}} = \frac{nq}{\sqrt{npq}}$$

$$= \sqrt{\frac{nq \cdot nq}{npq}} = \sqrt{\frac{nq}{p}}$$

Thus in the limit as $n \rightarrow \infty$, z takes the values from $-\infty$ to ∞

Hence the distribution of x will be a continuous distribution over the range $-\infty$ to ∞

Taking limits as $n \rightarrow \infty$ in (1) & (2)

using stirling's approxination to $\gamma!$ for large γ namely.

$$\lim_{\gamma \to \infty} \gamma! \simeq \sqrt{2\pi}\, e^{-\gamma}\, \gamma^{\gamma + 1/2}$$

Eve have in the limit as $n \rightarrow \infty$ and consequently $x \rightarrow \infty$

$$\lim P(x) = \lim \frac{n!}{x!(n-x)} p^x q^{n-x} \quad (5)$$

$$= \frac{\lim \sqrt{2\pi}\, e^{-n}\, n^{n+1/2} \cdot p^x q^{n-x}}{\sqrt{2\pi}\, e^{-x}\, x^{x+1/2} \sqrt{2\pi}\, e^{-(n-x)} (n-x)^{n-x+1/2}}$$

$$= \frac{\lim e^{-n}\, n^{n+1/2}\, p^x q^{n-x}}{(e^{-x-n+x}\, x^{x+1/2})(n-x)^{n-x+1/2} \sqrt{2\pi}}$$

$$= \frac{\lim n^{n+1/2}\, p^x q^{n-x}}{\sqrt{2\pi}\, x^{x+1/2} (n-x)^{n-x+1/2}}$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{npq}} \frac{(np)^{x+1/2}}{x^{x+1/2}} \frac{(nq)^{n-x+1/2}}{(n-x)^{n-x+1/2}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \left(\frac{np}{x}\right)^{x+1/2} \left(\frac{nq}{n-x}\right)^{n-x+1/2} \right] \rightarrow (3)$$

From (2), we get

$$x = np + z\sqrt{npq} \Rightarrow \frac{x}{np} = 1 + z\sqrt{\frac{q}{np}}$$

Further $n - x = n - np - z\sqrt{npq}$

$$\Rightarrow \frac{n-x}{np} = 1 - z\sqrt{\frac{p}{nq}}$$

Also, $dz = \frac{1}{\sqrt{npq}}\, dx$

Hence the Probability differential of the distribution of z in the limit is

$$dG(z) = g(z) \Rightarrow dz = \lim_{n \to \infty} \left(\frac{1}{\sqrt{2\pi}} \times \frac{1}{N} \right) dz \rightarrow (4)$$

where $N = \left(\frac{x}{np}\right)^{x+1/2} \left(\frac{n-x}{nq}\right)^{n-x+1/2}$

$$\log N = \log\left(\frac{x}{np}\right)^{x+1/2} + \log\left(\frac{n-x}{nq}\right)^{n-x+1/2} \quad (6)$$

$$= (x+1/2)\log\left(\frac{x}{np}\right) + (n-x+1/2)\log\left(\frac{n-x}{nq}\right)$$

$$= (np + z\sqrt{npq} + 1/2)\left\{z\cdot\sqrt{\frac{q}{np}} - \frac{1}{2}z^2\left(\frac{q}{np}\right) + \frac{1}{3}z^3\left(\frac{q}{np}\right)^{3}\right\}$$

$$+(np - z\sqrt{npq} + 1/2)\left\{-z\sqrt{\frac{p}{nq}} - \frac{1}{2}z^2\left(\frac{p}{nq}\right) - \frac{1}{3}z^3\left(\frac{p}{nq}\right)^{3/2}\cdots\right\}$$

$$= \frac{z^2}{2} + O(n^{-1/2}) \Rightarrow \frac{z^2}{2} \text{ as } n\to\infty$$

$\therefore$ $\lim_{n\to\infty} \log N = \frac{z^2}{2}$

$$\lim_{n\to\infty} N = e^{z^2/2}$$

Substitute these in (4)

$$dG(z) = g(z)\,dz = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \longrightarrow (5)$$

Hence the Probability function of z is

$$q(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty \longrightarrow (6)$$

* This is the Probability density function of the normal distribution mean and unit variance.

* If x is normal variate with mean μ and standard deviation σ then $z = x - \frac{\mu}{\sigma}$ is standard.

* Jacobian of transformation is $\frac{1}{\sigma}$. Hence substituting in (5), the P.d.f of a normal variate x with $E(x) = \mu$, $\text{var}(x) = \sigma^2$ is given by

$$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, & -\infty < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

chief characteristics of the normal distribution and normal Probability curve :-

The normal Probability curve with mean μ and standard deviation σ is given by the equation

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

and has the following Properties

(i) The curve is bell-shaped and symmetrical about the line $x = \mu$.

(ii) Mean, Median and mode of the distribution coincide.

(iii) As, x increases numerically, $f(x)$ decreases rapidly, the maximum Probability occuring at the Point $x = \mu$, and is given by

$$[P(x)]_{max} = \frac{1}{\sigma\sqrt{2\pi}}$$

(iv) $\beta_1 = 0$ and $\beta_2 = 3$

(v) $\mu_{2r+1} = 0 \quad (r = 0, 1, 2, \cdots)$ and $\mu_{2r} = (1)(3)(5)\cdots(2r-1)\sigma^{2r}$, $(r = 0, 1, 2, 3 \cdots)$

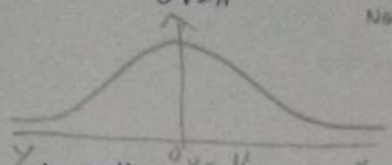
(vi) Since $f(x)$ being the Probability, can never be negative, no Portion of the curve, lies between the x-axis

(vii) Linear combination of independent normal variates is also a normal variates.

(viii) x-axis is an asympote to the curve.

(ii) ... of Inflexion of the curve are,

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



(x) Mean deviation about mean $= \sqrt{\frac{2}{\pi}}\sigma \simeq \frac{4}{5}\sigma$

(xi) Quartile are given by $Q_1 = \mu - 0.6745\sigma$;

$Q_3 = \mu + 0.6745\sigma$

(xii) $Q.D = \frac{Q_3 - Q_1}{2} \simeq \frac{2}{3}\sigma$ we have (approximate)

$Q.D : M.D : S.D : \frac{2}{3}\sigma ; \frac{4}{3}\sigma :: \frac{2}{3} : \frac{4}{5} : 1 =$

$Q.D : M.D : S.D :: 10 : 12 : 15$

(xiii) Area Probability

$P(\mu - \sigma < x < \mu + \sigma) = 0.6826$,

$P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544$;

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

The adjoint table gives the under the normal Probability curve for some important values of standard normal variable z.

Distance from the mean ordinates in terms of $\pm\sigma$	Area under the curve
$z = \pm 0.745$	50 % = 0.50
$z = \pm 1.00$	68.26 % = 0.6826
$z = \pm 1.96$	95 % = 0.95
$z = \pm 2.0$	95.44 % = 0.9544
$z = \pm 2.58$	99 % = 0.99
$z = \pm 3.0$	99.73 % = 0.9973

(xiv) If x and y are independent standard normal variates then it can be easily proved that,

$$U = X + Y$$

$$V = X - Y$$

are independently distributed, $U \sim N(0, 2)$ and $V \sim N(0, 2)$. We state (without Proof) the converse of this result which is due to D. Bernstein.

5m **Mode of normal distribution:**

Mode is the value of x for which $f(x)$ is maximum.

(i.e) mode is the solution of $f'(x) = 0$ and $f''(x) < 0$

For normal distribution with mean μ and standard deviation σ.

$$\log f(x) = c - \frac{1}{2\sigma^2}(x-\mu)^2$$

where, $c = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$ is a constant

$$\frac{1}{f(x)} \quad f'(x) = \frac{-1}{\sigma^2}(x-\mu)$$

$$\Rightarrow f'(x) = \frac{-1}{\sigma^2}(x-\mu) f(x) \quad \text{and} \quad \cancel{f''(x) = \frac{-1}{\sigma^2}}$$

$$f''(x) = \frac{-1}{\sigma^2}\left[1 \cdot f(x) + (x-\mu) f'(x)\right]$$

$$= \frac{-f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$$

$$f(x) = 0 \Rightarrow x - \mu = 0$$

$$\Rightarrow x = \mu$$

At Point $x = \mu$, we have

$$f''(x) = \frac{-1}{\sigma^2}\left[f(x)\right]_{x=\mu}$$

$$= \frac{-1}{\sigma^2}\frac{1}{\sigma\sqrt{2\pi}} < 0.$$

Hence $x < \mu$ is the mode, of the normal distribution.

(iv) Median of Normal distribution :-

If M is the median of normal distribution we have

(or) question $\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2} \Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = \frac{1}{2}$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = \frac{1}{2} \rightarrow (1)$$

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(-\frac{z^2}{2}\right) dz$$

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

From (1), we have

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

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = 0$$

$(i.e) \; \mu = M$

Hence for the normal distribution

Mean = median

(v) M.G.F of normal distribution :-

The m.g.f (about origin) is given by,

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

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

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

$$= e^{\mu t} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[z^2 - 2t\sigma z\right]\right\} dx$$

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

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

(11)

$$= e^{\mu t + t^2\sigma^2/2} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) du.$$

Hence, $\mu_x(t) = e^{\mu t + t^2\sigma^2/2}$.

(*) Moment of normal distribution:-

odd order moments about mean are given by,

(or) Question $\mu_{2n+1} = \int_{-\infty}^{\infty} (x-\mu)^{2n+1} f(x)\, dx$

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n+1} \exp\left(-\frac{z^2}{2}\right) dz.$$

$$\mu_{2n+1} = \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n+1} \exp\left(\frac{-z^2}{2}\right) dz = 0 \rightarrow (1)$$

since the integral $z^{2n+1} e^{-z^2/2}$ is an odd function.

(*) Even order moments about mean are given by

10m

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

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

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}}\, 2 \int_{0}^{\infty} z^{2n} e^{\left(-z^2/2\right)} dz$$

$$= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_{0}^{\infty} (2t)^n e^{-t} \frac{dt}{\sqrt{2t}}$$

$$\therefore \mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t}\, t^{(n+1/2)-1}\, dt$$

$$= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \sqrt{n+\tfrac{1}{2}}$$

changing 'n' into (n-1) we get,

$$\mu_{2n-2} = \frac{2^{n-1} \sigma^{2n-2}}{\sqrt{\pi}} \sqrt{n-\tfrac{1}{2}}$$

$$\frac{\mu_{2n}}{\mu_{2n-2}} = \frac{2\sigma^2 \sqrt{n+\tfrac{1}{2}}}{(n-\tfrac{1}{2})}$$

$$= 2\sigma^2 (n-\tfrac{1}{2})$$

$$\Rightarrow \mu_{2n} = \sigma^2 (2n-1) \mu_{2n-2} \rightarrow (2)$$

which since the recurrence relation for the moments of normal distribution from (2), we have,

$$\mu_{2n} = [(2n-1)\sigma^2][(2n-3)\sigma^2] \mu_{2n-4}$$

$$= [(2n-1)\sigma^2][(2n-3)\sigma^2][(2n-5)\sigma^2] \mu_{2n-6}$$

$$= [(2n-1)\sigma^2][(2n-3)\sigma^2][(2n-5)\sigma^2] \ldots [3\sigma^2][1\sigma^2]\mu_0$$

$$= 1.3.5 \ldots (2n-1)\sigma^{2n} \rightarrow (3)$$

From (1) & (2) we conclude that for the normal distribution all odd order moments about mean vanish and even order moments about mean are given by eqn (3)

A linear combination of independent normal variates is also a normal variate.

Soln:- Let x_i $(i = 1, 2, \ldots n)$

be n is independent normal variates with mean μ_i and variance σ_i^2 respectively.

Then $\mu_x(t) = a \rightarrow (1)$

The m.g.f of their linear combination $\sum_{i=1}^{n} a_i x_i$

where $a_1, a_2, \ldots a_n$ are constant is given by,

$$\mu_{\sum_{i=1}^{n} a_i x_i}(t) = \prod_{i=1}^{n} \mu_{a_i x_i}(t)$$

$$= \mu_{x_1}(a_1 t)\, \mu_{x_2}(a_2 t) \cdots \mu_{x_n}(a_n t) \rightarrow (2)$$

From (1) we have

$$\mu_{x_i}(a_i t) = e^{\mu_i a_i t + t^2 a_i^2 \frac{\sigma_i^2}{2}}$$

$$\therefore \mu_{\sum_{i=1}^{n} a_i x_i}(t) = \left[e^{\mu_1 a_1 t + t^2 a_1^2 \frac{\sigma_1^2}{2}} \times e^{\mu_2 a_2 t + t^2 a_2^2 \frac{\sigma_2^2}{2}} \times \cdots \right.$$

$$\left. \cdots \times e^{\mu_n a_n t + t^2 a_n^2 \frac{\sigma_n^2}{2}} \right]$$

$$= \left[\left(\sum_{i=1}^{n} a_i \mu_i \right) t + t^2 \left(\sum_{i=1}^{n} a_i^2 \frac{\sigma_i^2}{2} \right) \right]$$

which is the m.g.f of a normal variate with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$

Hence by the uniqueness theorem of m.g.f

$$\sum_{i=1}^{n} a_i x_i \sim N \left[\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right]$$

Points of inflexion of normal curve :-

At the Point of inflexion of the normal curve, we should have $f''(x) = 0$ and $f'''(x) \neq 0$

For normal curve, we have from (9.3)

$$f''(x) = \frac{-f(x)}{\sigma^2} \left\{ 1 - \frac{(x-\mu)^2}{\sigma^2} \right\}$$

$$f''(x) = 0 \Rightarrow \left[1 - \frac{(x-\mu)}{\sigma^2} \right] = 0$$

$$\Rightarrow x = \mu \pm 6$$

$x = \mu \pm 6$, $f'''(x) \neq 0$, Hence the Points of

Inflection of the normal curve are given by $x \pm \sigma$ and $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}}$ (i.e) they are equi distant (at a distance σ) from the mean.

Mean deviation about the mean for normal distribution :-

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

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

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

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

[$\because$ The integral $|z| e^{-z^2/2}$ is an even f'n of z]

Since in $[0, \infty]$; $|z| = z$; we have

$$\text{M.D} = \sqrt{\frac{2}{\pi}}\, \sigma \int_0^{\infty} e^{-z^2/2} dz$$

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

$$= \sqrt{\frac{2}{\pi}}\, \sigma \left| \frac{e^{-t}}{-1} \right|_0^{\infty}$$

$$= \sqrt{\frac{2}{\pi}}\, \sigma$$

$$= \frac{4}{5}\sigma \quad \text{(approximately)}$$

Area Property (Normal Probability integral) :-

If $x \sim N(\mu, \sigma^2)$ then the Probability that random value of x will lie between $x = \mu$ and $x = x_1$ is given by,

$$P(\mu < x < x_1) = \int_{\mu}^{x_1} f(x)\,dx$$

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

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

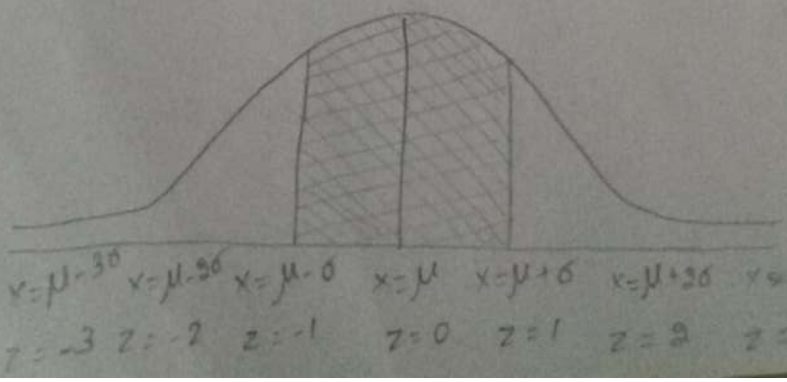
when $x = \mu$, $z = 0$ and $x = x_1$

$$z = \frac{x_1 - \mu}{\sigma} = z_1$$

$$P(\mu < x < x_1) = P(0 < z < z_1)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{z_1} e^{-z^2/2}\,dz = \int_{0}^{z_1} p(z)\,dz$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, is the Probability function of standard normal variate. The definite integral $\int_0^{z_1} \phi(z)\,dz$ is known as normal Probability integral and gives the area under standard normal curve between the co-ordinates at $z = 0$ and $z = z_1$, these areas have been tabulated for different value of z_1 at intervals of 0.01 in a table given at the end of the chapter.



In Particular the Probability that a random (16)
vaule of x lies in the interval $(\mu-\sigma, \mu+\sigma)$
is given by

$$P(\mu-\sigma < x < \mu+\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\,dx$$

$$P(-1 < z < 1) = \int_{-1}^{1} \phi(z)\,dz$$

$$= 2\int_{0}^{1} \phi(z)\,dz$$

$$= 2 \times 0.3413$$

$$= 0.6826$$

similarly

$$P(\mu-2\sigma < x < \mu+2\sigma) = P(-2 < z < 2)$$

$$= \int_{-2}^{2} \phi(z)\,dz = 2\int_{0}^{2} \phi(z)\,dz$$

$$= 2 \times 0.4772 = 0.9544$$

and

$$P(\mu-3\sigma < x < \mu+3\sigma) = P(-3 < z < 3)$$

$$= \int_{-3}^{3} \phi(z)\,dz = 2\int_{0}^{3} \phi(z)\,dz$$

$$= 2 \times 0.49865 = 0.9973$$

Thus the Probability that a normal variates x lies between the range $\mu \pm 3\sigma$ is given by

$$P[|x-\mu| > 3\sigma] = P[|z| > 3]$$

$$= 1 - P(|z| < 3)$$

$$= 1 - P(-3 \leq z \leq 3)$$

$$= 0.0027$$

Thus in all Probability, we should expect a normal variate to be within the range $\mu \pm 3\sigma$ though theoretically it may range from $-\infty$ to ∞.

Problem -1 (11)

Obtain the equation of the normal curve that may be fitted to the following data.

class	60-65	65-70	70-75	75-80	80-85	85-90	90-95	95-100
Frequency	3	21	150	335	326	135	26	4

soln:- For the given data $N = 1000$, $\mu = 79.945$ and $\sigma = 5.545$

Hence the equation of the normal curve fitted to the given data is computation of theoretical normal frequency

class interval	lower class boundary x	$z = \frac{x-\mu}{\sigma}$	$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\,du$	$A\phi(z) = \phi z_H - \phi z$	expected frequency
Below 60	$-\infty$	$-\infty$	0	0.00112	$0.12 \cong 0$
60-65	60	-3.663	0.000112	0.002914	$2.914 \cong 3$
65-70	65	-2.745	0.003026	0.031044	$31.044 \cong 31$
70-75	70	-1.826	0.034070	0.147870	$147.870 \cong 11.8$
75-80	75	-0.908	0.181940	0.322050	$322.050 \cong 3.02$
80-85	80	0.10	0.503990	0.919300	$319.300 \cong 315$
85-90	85	0.928	0.823290	0.164072	$144.072 \cong 144$
90-95	90	1.487	0.967362	0.29792	$29.772 \cong 30$
95-100	95	2.675	0.997154	0.002733	$2.733 \cong 3$
100 and over	100	3.683	0.999887		1000

2. X is normally distributed and the mean of x is 2 and S.D is 4. a) Find the Probability of following

(i) $x \geq 20$, (ii) $x \leq 20$, (iii) $0 \leq x \leq 12$.

b) Find x', when $P(x > x') = 0.24$.

c) Find x_0' and x_1' when $P(x_0' < x < x_1') = 0.50$ and $P(x > x') = 0.25$.

Soln:- a) we have $\mu = 12$, $\sigma = 4$. $x \sim N(12, 4)$

i) $P(x \geq 20)$

when $x = 20$, $z = \frac{x-\mu}{\sigma} = \frac{20-12}{4}$, $z = 2$

e that may

95 95-100

4

nd $\sigma = 5.5745$

ed to the

al frequency

expected frequency

0.12 ≅ 0

2.914 ≅ 3

31.044 ≅ 31

47.870 ≅ 11.8

322.050 ≅ 3.02

19.300 ≅ 319

4.072 ≅ 144

7.772 ≅ 30

733 ≅ 3

1000

x is 2.

g

).50

a) $P(x \geq 20) = P(z \geq 2)$

$= 0.5 - P(0 \leq z \leq 2)$

$= 0.5 - 0.4772$

$= 0.0228$

ii) $P(x \leq 20) = 1 - P(x \geq 20)$

$= 1 - 0.0228$

$= 0.9722$

iii) $P(0 \leq x \leq 12) = P(-3 \leq z \leq 0)$

$= P(0 \leq z \leq 3)$

$= 0.49865$

b) when $x = x'$ $\quad z = \frac{x' - 12}{4} = z_1$

Then, $P(x > x') = 0.24$

$P(z > z') = 0.24$

$P(0 < z < z') = 0.26$

c) Given $P(x_0' \leq x \leq x_1') = 0.50$ and

$P(x > x_1') = 0.25$

$x = x_1'$; $\quad z = \frac{x_1' - 12}{4} = z_1$

$x = x_0'$; $\quad z = \frac{x_0' - 12}{4} = -z_1$

$P(z > z_1) = 0.25$

$P(0 < z < z_1) = 0.25$

Hence, $\frac{x_1' - 12}{4} = 0.67$

$x_1' = 0.67 \times 4 + 12$

$x_1' = 14.68$.

and

$\frac{x_0' - 12}{4} = 0.67$

$x_0' = 12.4 \times 0.67$

$x_0' = 9.32$

3. x is a normal variance with mean 30 and S.D 5. Find the Probability that, (i) $26 \leq x \leq 40$ (ii) $x \geq 45$ (iii) $|x-30| > 5$ (4)

Soln:- Hence, $\mu = 30$, $\sigma = 5$

i) when $x = 26$, $z = \frac{x-\mu}{\sigma} = \frac{26-30}{5}$, $z = -0.8$ $\quad x = 40$

$$z = \frac{40-30}{5} = 2$$

$$\therefore P(26 \leq x \leq 40) = P(-0.8 \leq z \leq 2)$$

$$= P(-0.8 \leq z \leq 6) + P(0 \leq z \leq 2)$$

$$= P(0 \leq z \leq 0.8) + P(0 \leq z \leq 2)$$

$$= 0.288 + 0.4772$$

$$= 0.7653$$

ii) when $x = 45$, $z = \frac{45-30}{5} = 3$

$$P(x \geq 45) = P(z > 3)$$

$$= 0.5 - 0.49865$$

$$= 0.00135$$

iii) $P[|x-30| \leq 5] = P(25 \leq x \leq 35)$

$$= P(-1 \leq z \leq 1)$$

$$= 2P(0 \leq z \leq 1)$$

$$= 2 \times 0.3413$$

$$= 0.6826$$

$$\therefore P|x-30| > 5$$

$$= 1 - P(|x-30| \leq 5)$$

$$= 1 - 0.6826$$

$$= 0.3174$$

4. The mean yield for one acre plot is 662 bilos with a standard deviation 32 kilos. Assuming normal distribution (20) How many one-acre plots in a batch of 1000 plots would you expect to have yield.

i) over 700 bilos (ii) below 650 bilos (iii) what is the lowest yield of the best 100 plots.

Sol:- If the random variable x denotes the yield (in bilos) for one acre plot, where $\mu = 662$ and $\sigma = 32$.

i) The Probability that a plot has a yield over 700 bilos is given by $P(x > 700) = P(z > 1.19)$

$z = \frac{700 - 662}{32}$

$= 0.5 - 0.3830$

$= 0.1170$

$z = \frac{700 - 662}{32} = \frac{38}{32} = \frac{19}{16} = 1.19$

Hence in a batch of 1000 plots, the expected no. of plots with yield over 700 bilos is $1000 \times 0.117 = 117$

ii) Required No. of Plots with yield below 650 bilos is given by:-

$1000 \times P(x < 650) = 1000 \times P(z < -0.38)$

$= 1000 \times P(z > 0.38)$

$= 1000 \times [0.5 - P(0 \le z \le 0.38)]$

$= 1000 \times [0.5 - 0.1480]$

$= 1000 \times 0.352 = 352$

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

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

when $x = x_1$, $z = \frac{x_1 - \mu}{\sigma} = \frac{x_1 - 662}{32} = z_1 \rightarrow (1)$

Such that,

$P(z > z_1) = 0.1$

$P(0 \le z \le z_1) = 0.4$

$\rightarrow z_1 = 1.28$

Substituting in ①,

$x_1 = 662 + 32z_1$

$= 662 + 32 \times 1.28 = 662 + 40.96$

$= 702.96$

Hence the best 100 Plots have yield over 702.96 kilos.

Moment Generation function:-

The moment generating function. (m.g.f) of a random variable x (about) the origin having the Probability function $f(x)$ is given by.

$M_x(t) = E(e^{tx})$

$= \int e^{tx} f(x)\, dx$, for continuous.

$\sum_x e^{tx} f(x)$ for discrete.

Property - 1

$M_{cx}(t) = M_x(ct)$ $c \to a$, constant.

Proof:-

$LHS = M_{cx}(t)$

$= E\left[e^{tcx}\right] \to ①$

$RHS = M_x(ct)$

$= E\left[e^{ctx}\right] \to ②$

From ① & ②, we get

$LHS = RHS \Rightarrow M_{cx}(t) = M_x(ct)$

Property - ②

The moment generating function of the sum of a number of independent random variables is equal to the Product of the respective moment generating function.

Symbolically if $x_1, x_2, \cdots x_n$ are independent random variables, then the moment Generating function of their sum $x_1 + x_2 + \cdots + x_n$ is given by.

$M_{x_1 + x_2 + \cdots + x_n}(t) = M_{x_1}(t) \cdot M_{x_2}(t) \cdots M_{x_n}(t)$

Proof:- By definition,

$$M_{x_1+x_2+\ldots+x_n}(t) = E\left[e^{t(x_1+x_2+\ldots+x_n)}\right]$$

$$= E\left[e^{tx_1} \cdot e^{tx_2} \cdots e^{tx_n}\right]$$

$$= E\left[e^{tx_1}\right) \cdot E\left[e^{tx_2}\right] \ldots E\left[e^{tx_n}\right]$$

$\because (x_1, x_2, \ldots x_n$ are independent)

$$= M_{x_1}(t) \cdot M_{x_2}(t) \ldots\ldots M_{x_n}(t)$$

Property -3

Effect of change of origin and scale on M.G.F

Soln:- Let us transform x to u by chaing both the origin and scale in x as follows.

$$U = \frac{x-a}{h}, \text{ where } a \text{ and } h \text{ are constants.}$$

M.G.F of u (about origin) is Given,

$$M_u(t) = E(e^{tu})$$

$$= E\left[e^{t\left(\frac{x-a}{h}\right)}\right)$$

$$= E\left[e^{tx/h} \cdot e^{-at/h}\right)$$

$$= e^{-at/h}\left[E\left[e^{tx/h}\right]\right]$$

$$= e^{-at/h} M_x\left(t/h\right)$$

where $M_x(t)$ is the M.g.f of x about origin

In Particular, If we taken,

$a = E(x) = \mu$ and $h = \sigma x = \sigma$

Then $u = \frac{x - E(x)}{\sigma x} = \frac{x-\mu}{\sigma} = z$ is known as a standard ~~deviate~~ variate.

Thus the M.g.f of a standard variate z is given by

$$M_z(t) = e^{-\mu t/\sigma} M_x\left(\frac{t}{\sigma}\right)$$

Cumulants :-

Cumulants generating function $K(t)$ is defined as $K_x(t) = \log_e M_x(t)$ Provided the right hand side can be Expanded as a convergent series in Powers of t.

Thus, $K_x(t) = K_1(t) + K_2 \frac{t^2}{2!} + \ldots + K_r \frac{t^r}{r!}$

$= \log M_x(t)$

$= \log E\left[e^{tx}\right]$

$= \log E\left[1 + tx + \frac{t^2x^2}{2!} + \ldots + \frac{t^r x^r}{r!}\right]$

Rectangular or uniform distribution on $[a, b]$

$$f[x; a, b] = \frac{1}{b-a}, \quad a < x < b$$

$$F(x) = \begin{cases} \frac{x-a}{b-a} & a < x < b \\ x \geq b \end{cases}$$

$$\frac{d}{dx} F(x) = f(x)$$

$X \sim U[a, b]$

$X \sim R[a, b]$

Rectangular distribution function of $[-a, a]$

$$f(x) = \frac{1}{a-(-a)}, \quad -a < x < a$$

$$= \frac{1}{2a} \quad -a < x < a$$

$$\mu_r' = \int x^r f(x)\, dx$$

$$\mu_r' = \int_a^b x^r \frac{1}{b-a}\, dx$$

$$= \frac{1}{b-a} \int_a^b x^r\, dx$$

Mean $= \mu_1 \Rightarrow \mu_1' = \frac{1}{b-a}\int_a^b x\,dx$

$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right]$

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

$= \frac{a+b}{2}$

$\mu_2' = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b$

$= \frac{1}{b-a}\left[\frac{b^3-a^3}{3}\right]$

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

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

Variance $\mu_2 = \mu_2' - (\mu_1')^2$

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

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

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

$M_X(t) = E(e^{tx}) = \int_a^b e^{tx}\cdot f(x)\,dx$

$= \int_a^b e^{tx}\cdot\frac{1}{b-a}\,dx$

$= \frac{1}{b-a}\int_a^b e^{tx}\,dx = \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_a^b$

$= \frac{1}{b-a}\left[\frac{e^{tb}-e^{ta}}{t}\right]$

Characteristic function:-

$$\phi_x(t) = \frac{1}{b-a}\int_a^b e^{itx}\,dx$$

$$= \frac{1}{b-a}\left[\frac{e^{ibt} - e^{iat}}{it}\right], \quad t \neq 0$$

Mean deviation about mean.

$$\frac{\Sigma f|x-\bar{x}|}{\Sigma f} \qquad \eta = E(|x - \text{mean}|)$$

$$= \int_a^b |x - \text{mean}|\, f(x)\,dx$$

$$= \frac{1}{b-a}\int_a^b |x - \text{mean}|\,dx$$

$$= \frac{1}{b-a}\int_a^b \left|x - \left(\frac{a+b}{2}\right)\right| dx$$

$\left(\frac{b}{2} - \frac{a}{2}\right)$ $\qquad t = x - \frac{a+b}{2}$

$$= \frac{1}{b-a}\int |t|\,dt, \qquad dt = dx$$

$-\left(\frac{b}{2} - \frac{a}{2}\right)$ $\qquad x = a, \; t = \frac{a}{2} - \frac{b}{2}$

$\qquad\qquad\qquad\qquad x = b, \; t = \frac{b}{2} - \frac{a}{2}$

$$\neq f = \frac{1}{b-a}\,2\int_0^{\frac{b}{2}-\frac{a}{2}} t\,dt$$

$$= \frac{2}{b-a}\left[\frac{t^2}{2}\right]_0^{\frac{b}{2}-\frac{a}{2}}$$

$$= 2.$$

③rd Property cumulantes continues.

$$= \log\left[1 + t\,E(x) + \frac{t^2x^2}{2!} + \ldots + \frac{t^rx^r}{r!} + \ldots\right]$$

$$= \log\left[1 + t\,E(x) + \frac{t^2}{2!}E(x^2) + \ldots + \frac{t^r}{r!}E(x^r) + \ldots\right]$$

$$= \log\left[1 + t\mu_1' + \mu_2'\frac{t^2}{2!} + \mu_3'\frac{t^3}{3!} + \ldots + \mu_r'\frac{t^r}{r!} + \ldots\right]$$

$\mu_r' = E(x)^r$

$\mu_1' = E(x), \; \mu_2' = E(x^2)$

where μ_r = coefficient of $\frac{t^r}{r!}$ in $K_x(t)$ is called the rth cumulants. (26)

Hence,

$$K_1 t + K_2 \frac{t^2}{2!} + K_3 \frac{t^3}{3!} + K_4 \frac{t^4}{4!} + \cdots$$

$$= \left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \cdots\right) - \frac{1}{2}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \cdots\right)^2 + \frac{1}{3}\left[\mu_1' t + \mu_2' \frac{t^2}{2!} + \cdots\right]^3 - \frac{1}{4}\left[\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \cdots\right]^4$$

Comparing the co-efficients of like Powers of "t" on both sides, we get the relationship between the moments and cumulants.

Hence $\boxed{K_1 = \mu_1' = \text{Mean}}$ (X)

$$\frac{K_2}{2!} = \frac{\mu_2'}{2!} - \frac{\mu_1'^2}{2!}$$

$$\Rightarrow K_2 = \mu_2' - \mu_1'^2 = \mu_2$$

$$\frac{K_3}{3!} = \frac{\mu_3'}{3!} - \frac{1}{2}\frac{2\mu_1'\mu_2'}{2!} + \frac{\mu_1'^3}{3!}$$

$$\Rightarrow K_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \mu_3$$

Also, $$\frac{K_4}{4!} = \frac{\mu_4'}{4} - \frac{1}{2}\left[\frac{\mu_2'^2}{4} + \frac{2\mu_1'\mu_3'}{3!}\right] + \frac{1}{3}\frac{3\mu_1'^2\mu_2'}{2} - \frac{\mu_1'^4}{24}$$

$$\Rightarrow K_4 = \mu_4' - 3\mu_2'^2 - 4\mu_1'\mu_3' + 12\mu_1'^2\mu_2' - 6\mu_1'^4$$

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

$$= \mu_4 - 3(\mu_2' - \mu_1'^2)^2$$

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

$$\mu_4 = K_4 + 3K_2^2 \quad (\because \mu_2 = K_2)$$

Hence we have obtained.

Mean $= k_1$

$\mu_2 = k_2$ = variance

$\mu_3 = k_3$

$\mu_4 = k_4 + 3k_2^2$.

Properties of Cumulants :-

Property ① Additive Property of cumulants :-

The rth cumulants of the sum of the indep random variables is equal to the sum of the rth cumulants of the individual variables. Symbolically

$$k_r(x_1 + x_2 + \dots + x_n) = k_r(x_1) + k_r(x_2) + \dots + k_r(x_n)$$

where x_i ; $i = 1, 2, \dots n$ are independent r.v.s

Proof :-

Since x_i's are independent

$$M_{x_1 + x_2 + x_3 + \dots + x_n}(t) = M_{x_1}(t)\, M_{x_2}(t) \dots M_{x_n}(t)$$

Taking logarithm of each side

$$\log\left[M_{x_1 + x_2 + \dots + x_n}(t)\right] = \log M_{x_1}(t)\, M_{x_2}(t) \dots M_{x_n}(t)$$

$$K_{x_1 + x_2 + x_3 + \dots + x_n}(t) = \log M_{x_1}(t) + \log M_{x_2}(t) + \dots + \log M_{x_n}(t)$$

$$= K_{x_1}(t) + K_{x_2}(t) + K \dots + K_{x_n}(t)$$

Differentiating both side w.r to "t" r times and Putting $t = 0$, we get,

$$\left[\frac{d^r}{dt^r}\left(K_{x_1 + x_2 + \dots + x_n}(t)\right)\right]_{t=0} =$$

$$= \left[\frac{d^r}{dt^r} K_{x_1}(t)\right]_{t=0} + \left[\frac{d^r}{dt^r} K_{x_2}(t)\right]_{t=0} + \dots + \left[\frac{d^r}{dt^r} K_{x_n}(t)\right]_{t=0}$$

$\left[\because K_r = \left[\frac{d^r}{dt^r} K_x(t)\right]_{t=0}\right]$

$\therefore K_r(x_1 + x_2 + \ldots + x_n) = K_r(x_1) + K_r(x_2) + \ldots + K_r(x_n)$

Property – ② :-

Effect of change of origin and scale on cumulants.

Proof:- If we take $u = \frac{x-a}{h}$

then, $M_u(t) = E(e^{tu}) = E\left(e^{\left(\frac{x-a}{h}\right)t}\right]$

$= E\left[e^{\frac{tx}{h}} \cdot e^{-at/h}\right]$

$= e^{-\frac{at}{h}} E\left[e^{\frac{tx}{h}}\right]$

$= e^{-at/h} M_x(t/h)$

$K_u(t) = \log M_u(t)$

$= \log\left[e^{-at/h} \cdot M_x(t/h)\right]$

$= \log_e\left[-at/h + \log M_x(t/h)\right]$

$= -\frac{at}{h} + K_x(t/h)$

$K_1' + K_2' \frac{t^2}{2!} + \ldots + K_r' \frac{t^r}{r!} + \ldots$

$= \frac{-at}{h} + K_1\left(\frac{t}{h}\right) + K_2 \frac{(t/h)^2}{2!} + \ldots$

$+ \frac{K_r (t/h)^r}{r!} + \ldots$

where K_r' and K_r are the rth cumulants of u and x respectively.

Comparing coefficients.

we get, $K_1' = \frac{K_1 - a}{h}$ and $K_r' = \frac{K_r}{h^r}$; $r = 2, 3 \ldots$

Thus we see that except the first cumulant, all cumulants are independent of change of origin. But the cumulants are not invariant of change of scale as the rth cumulant of u is $\left(\frac{1}{h^r}\right)$ times the rth cumulant of distribution of x.

Problem ①:-

Let the r.v. x assume the value "r" with the Probability law $P(x=r) = q^{r-1}.p$; $r = 1, 2, 3 \ldots$

Find the m.g.f of x and hence its mean and variance.

soln:- $M_x(t) = E(e^{tx}) = \sum_{r=1}^{\infty} e^{tr} q^{r-1}.p$

$$= \frac{p}{q} \sum_{r=1}^{\infty} (qe^t)^r$$

$$= \frac{p}{q} qe^t \sum_{r=1}^{\infty} (qe^t)^{r-1}$$

$$= pe^t \{1 + qe^t + (qe^t)^2 + \ldots\}$$

$$= \frac{pe^t}{1-qe^t}.$$

Diff w.r to "t". Then we have

$$M_x'(t) = \frac{pe^t}{(1-qe^t)^2}$$

$$M_x''(t) = pe^t \frac{(1+qe^t)}{(1-qe^t)^3}$$

$$\therefore \mu_1' \text{ (about origin)} = M_x'(0) = \frac{p}{(1-q)^2} = \frac{1}{p}$$

$$\mu_2' \text{ (about origin)} = M_x''(0) = \frac{p(1+q)}{(1-q)^3} = 1 + \frac{q}{p^2}$$

Hence mean = μ_1' (about origin)

$= 1/p$

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

$= \frac{1+q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}$

Problem - 2

The Probability density function of the random variable x follows the Probability law.

$$P(x) = \frac{1}{2\theta} \cdot e^{-\frac{|x-\theta|}{\theta}} \; ; \; -\infty < x < \infty$$

Find M.G.F of x. Hence or otherwise find $E(x)$ and $V(x)$.

Soln:- The moment generating function of x is,

$$M_x(t) = E(e^{tx})$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\theta} e^{-\frac{|x-\theta|}{\theta}} e^{tx} \, dx$$

$$= \int_{-\infty}^{\theta} \frac{1}{2\theta} e^{-\frac{\theta - x}{\theta}} \cdot e^{tx} \, dx + \int_{\theta}^{\infty} \frac{1}{2\theta} e^{-\frac{x-\theta}{\theta}} \cdot e^{tx} \, dx$$

[For $x \in (-\infty, \theta)$, $x - \theta < \theta \Rightarrow \theta - x > 0$.

$|x-\theta| = \theta - x \; \forall \; x \in (-\infty, \theta)$

similarly $|x-\theta| = x - \theta \; \forall \; x \in (\theta, \infty)$]

$$M_x(t) = \frac{e^{-1}}{2\theta} \int_{-\infty}^{\theta} e^{x(t+1/\theta)} \, dx + \frac{e}{2\theta} \int_{\theta}^{\infty} e^{-x\left[\frac{1}{\theta} - t\right]} \, dx$$

$$= \frac{e^{-1}}{2\theta} \cdot \frac{1}{t+1/\theta} e^{\theta(t+1/\theta)} + \frac{e}{2\theta} \cdot \frac{1}{1/\theta - t} e^{-\theta\left(\frac{1}{\theta} - t\right)}$$

$$= \frac{e^{\theta t}}{2(\theta t + 1)} + \frac{e^{\theta t}}{2(1 - \theta t)}$$

$$= \frac{e^{\theta t}}{1 - \theta^2 t^2}$$

$$= e^{\theta t} (1 - \theta^2 t^2)^{-1}$$

$$= \left(1 + \theta t + \frac{\theta^2 t^2}{2!} + \cdots\right)\left(1 + \theta^2 t^2 + \theta^4 t^4 + \cdots\right)$$

$$= 1 + \theta t + \frac{3\theta^2 t^2}{2!} + \cdots$$

$E(x) = \mu_1' =$ co-efficient of t in $M_x(t) = \theta$

$\mu_2' =$ co-eff of $\frac{t^2}{2!}$ in $M_x(t) = 3\theta^2$

Hence, $\text{var}(x) = \mu_2' - \mu_1'^2 = 3\theta^2 - \theta^2$

$= 2\theta^2$.

Problem ③ :-

If the moments of variate x are defined by $E(x^r) = 0.6$, $r = 1, 2, 3 \ldots$ show that $P(x=0) = 0.4$, $P(x=1) = 0.6$, $P(x \geq 2) = 0$.

soln:- The m.g.f of variate x is

$$M_x(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r' = 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!}(0.6)$$

$$= 0.4 + 0.6 \sum_{r=0}^{\infty} \frac{t^r}{r!}$$

$$= 0.4 + 0.6\, e^t \longrightarrow (1)$$

But,

$$M_x(t) = E(e^{tx})$$

$$= \sum_{x=0}^{\infty} e^{tx} P(x=x)$$

$$= P(x=0) + e^t \cdot P(x=1) + \sum_{x=2}^{\infty} e^{tx} P(x=2) \longrightarrow (2)$$

From (1) & (2) we get

$$P(x=0) = 0.4, \quad P(x=1) = 0.6, \quad P(x \geq 2) = 0.$$

Problem ④ :-

Find the m.g.f of the random variable whose moment are $\mu_r' = (r+1)!\, 2^r$.

soln:- The m.g.f is given by

$$M_x(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r'$$

$$= \sum_{r=0}^{\infty} \frac{t^r}{r!} (r+1)! 2^r$$

$$= \sum_{r=0}^{\infty} (r+1)(2t)^r$$

$$= 1 + 2(2t) + 3(2t^2) + \cdots$$

$$= (1-2t)^{-2}$$

characteristic function :-

$$\phi_x(t) = E(e^{itx})$$

$$= \int e^{itx} f(x)\,dx \quad ; \text{ for continuous}$$

$$= \sum_x e^{itx} P(x) \quad ; \text{ for discrete}$$

Formulae :-

$$\sqrt{k} = (k-1)\sqrt{k-1}$$

Theorem :-

$$G_Z(z) = P(Z \le z)$$

$$= 1 - P(Z > z)$$

$$= 1 - P(\min[(x_1, x_2, \cdots, x_n) > z]$$

$$= 1 - P(x_i^0 > z)$$

$$= 1 - \prod_{i=1}^{n} P(x_i^0 > z)$$

$$= 1 - \prod_{i=1}^{n} [1 - P(x_i^0 \le z)]$$

$$= 1 - \prod_{i=1}^{n} [1 - F_{x_i^0}(z)]$$

$$= 1 - \prod_{i=1}^{n} [1 - (1 - e^{-\theta z})]$$

$$G_Z(z) = 1 - e^{-\sum_{i=1}^{n} \theta_i^0 z}$$

$$g_Z(z) = \frac{d}{dz} G_Z(z)$$

$$= -\left[-\Sigma \theta_i\right] e^{-\Sigma \theta_i z}$$

$$= \begin{cases} \sum_{i=1}^{n} \theta_i \; e^{-\sum_{i=1}^{n} \theta_i z}, & z > 0 \\ 0 \; ; \text{ otherwise.} \end{cases}$$

Unit - I - Finish.

UNIT-4

Example : 14.1

Soln:-

$n = 9000, \quad x = 3,240$

P = Probability of getting 3 or 4

$$P = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \qquad Q = 1 - P = 1 - \frac{1}{3}$$

$$Q = \frac{2}{3}$$

$$Z = \frac{x - np}{\sqrt{nQP}} \sim N(0, 1)$$

Then $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$, for every x and hence $f(x)$ defined in (9·17) is a *p.d.f.* of X.

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

Moments. The rth moment about origin is given by :

$$\begin{aligned} \mu_r' &= E(X^r) = E(e^{rY}) && [\because Y = \log X \Rightarrow X = e^Y] \\ &= M_Y(r) && (m.g.f.\ of\ Y,\ r\ being\ the\ parameter) \\ &= \exp(\mu r + \tfrac{1}{2} r^2 \sigma^2) && [\because Y \sim N(\mu, \sigma^2)] \end{aligned} \quad \ldots (9·18)$$

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

$$\mu_r' = E(X^r) = \exp(r \log \alpha + \tfrac{1}{2} r^2 \sigma^2) = \alpha^r \exp(r^2 \sigma^2/2) \quad \ldots (9·18a)$$

$\therefore$ Mean $= \mu_1' = \alpha e^{\sigma^2/2}$ and $\mu_2 = \mu_2' - \mu_1'^2 = \alpha^2 e^{\sigma^2}(e^{\sigma^2} - 1)$

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

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

unit- II

9·3. RECTANGULAR (OR UNIFORM) DISTRIBUTION

Definition. *A random variable X is said to have a continuous rectangular (uniform) distribution over an interval (a, b), i.e., $(-\infty < a < b < \infty)$, if its p.d.f. is given by :*

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases} \quad \ldots (9·19)$$

Remarks 1. a and b, $(a < b)$ are the two parameters of the distribution. The distributin is called uniform distribution on (a, b) since it assumes a constant (uniform) value for all x in (a, b).

2. The distribution is also known as rectangular distribution, since the curve $y = f(x)$ describes a rectangle over the x-axis and between the ordinates at $x = a$ and $x = b$.

3. A uniform or rectangular variate X on the interval (a, b) is written as : $X \sim U[a, b]$ or $X \sim R[a, b]$.

4. The cumulative distribution function $F(x)$ is given by :

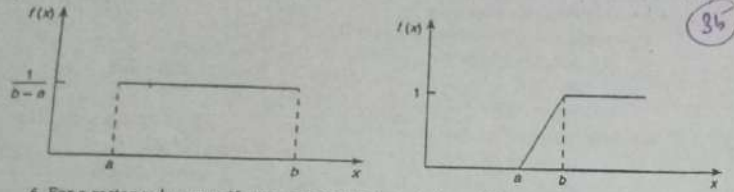
$$F(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \ge b \end{cases} \quad \ldots (9·19a)$$

Since $F(x)$ is not continuous at $x = a$ and $x = b$, it is not differentiable at these points. Thus $\frac{d}{dx} F(x) = f(x) = \frac{1}{b-a} \neq 0$, exists everywhere except at the points $x = a$ and $x = b$ and consequently *p.d.f.* $f(x)$ is given by (9·19).

Exp(x) discrete

$\int x f(x) dx$ Continuous

5. The graphs of uniform p.d.f. $f(x)$ and the corresponding distribution function $F(x)$ are given below.



6. For a rectangular or uniform variate X in $(-a, a)$, the p.d.f. is given by :

$$f(x) = \begin{cases} \frac{1}{2a}, & -a < x < a \\ 0, & \text{otherwise} \end{cases} \quad \ldots (9\cdot19b)$$

9·3·1. Moments of Rectangular Distribution. Let $X \sim U[a, b]$.

$$\mu'_r = \int_a^b x^r f(x)\,dx = \frac{1}{b-a}\int_a^b x^r\,dx = \frac{1}{b-a}\left(\frac{b^{r+1} - a^{r+1}}{r+1}\right) \quad \ldots (9\cdot20)$$

In particular

$$\text{Mean} = \mu_1' = \frac{1}{b-a}\left(\frac{b^2 - a^2}{2}\right) = \frac{b+a}{2} \quad \ldots (9\cdot20a)$$

and $$\mu_2' = \frac{1}{b-a}\left(\frac{b^3 - a^3}{3}\right) = \frac{1}{3}(b^2 + ab + a^2)$$

$$\therefore \text{Variance} = \mu_2' - \mu_1'^2 = \frac{1}{3}(b^2 + ab + a^2) - \left\{\frac{1}{2}(b+a)\right\}^2 = \frac{1}{12}(b-a)^2 \quad \ldots (9\cdot20b)$$

9·3·2. M.G.F. of Rectangular Distribution is given by :

$$M_X(t) = \int_a^b e^{tx} f(x)\,dx = \int_a^b \frac{e^{tx}}{b-a}\,dx = \frac{e^{bt} - e^{at}}{t(b-a)},\ t \neq 0 \quad \ldots (9\cdot20c)$$

9·3·3. Characteristic Function of Rectangular Distribution is given by :

$$\phi_X(t) = \int_a^b e^{itx}\,dx = \frac{e^{ibt} - e^{iat}}{it(b-a)},\ t \neq 0. \quad \ldots (9\cdot20d)$$

9·3·4. Mean Deviation about Mean, η of Rectangular Distribution is given by :

$$\eta = E|X - \text{Mean}| = \int_a^b |x - \text{Mean}|\,f(x)\,dx$$

$$= \frac{1}{b-a}\int_a^b \left|x - \frac{a+b}{2}\right| dx = \frac{1}{b-a}\int_{-(b-a)/2}^{(b-a)/2} |t|\,dt, \text{ where } t = x - \frac{a+b}{2}$$

$$= \frac{1}{b-a}\cdot 2\int_0^{(b-a)/2} t\,dt = \frac{b-a}{4} \quad \ldots (9\cdot20e)$$

Example 9·21. *If X is uniformly distributed with mean 1 and variance $\frac{4}{3}$ find $P(X < 0)$.*

Solution. Let $X \sim U[a, b]$, so that $p(x) = \frac{1}{b-a}$, $a < x < b$. We are given :

$$\text{Mean} = \frac{1}{2}(b+a) = 1 \implies b + a = 2 \text{ and } \text{Var}(X) = \frac{1}{12}(b-a)^2 = \frac{4}{3} \implies b - a = \pm 4.$$

Solving, we get $a = -1$ and $b = 3$; $(a < b)$. $\therefore\ p(x) = \frac{1}{4}$; $-1 < x < 3$

$$P(X < 0) = \int_{-1}^{0} p(x)\,dx = \frac{1}{4}\Big|x\Big|_{-1}^{0} = \frac{1}{4}.$$

Example 9·22. *Subway trains on a certain line run every half hour between mid-night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least twenty minutes ?*

Solution. Let the r.v. X denote the waiting time (in minutes) for the next train. Under the assumption that a man arrives at the station at random, X is distributed uniformly on (0, 30), with p.d.f.,

$$\therefore\ f(x) = \begin{cases} \frac{1}{30}, & 0 < x < 30 \\ 0, & \text{otherwise} \end{cases}$$

The probability that he has to wait at least 20 minutes is given by :

$$P(X \geq 20) = \int_{20}^{30} f(x)\,dx = \frac{1}{30}\int_{20}^{30} 1.\,dx = \frac{1}{30}(30 - 20) = \frac{1}{3}.$$

Example 9·23. *If X has a uniform distribution in $[0, 1]$, find the distribution (p.d.f.) of $-2 \log X$. Identify the distribution also.*

Solution. Let $Y = -2 \log X$. Then the distribution function G of Y is given by :

$$G_Y(y) = P(Y \leq y) = P(-2 \log X \leq y) = P(\log X \geq -y/2) = P(X \geq e^{-y/2})$$

$$= 1 - P(X \leq e^{-y/2}) = 1 - \int_0^{e^{-y/2}} f(x)\,dx = 1 - \int_0^{e^{-y/2}} 1.\,dx = 1 - e^{-y/2}$$

$$\therefore\ g_Y(y) = \frac{d}{dy}G(y) = \frac{1}{2}e^{-y/2},\ 0 < y < \infty \quad \ldots (*)$$

[$\because$ as X ranges in (0, 1), $Y = -2 \log X$ ranges from 0 to ∞]

Remark. This example illustrates that if $X \sim U[0, 1]$, then $Y = -2 \log X$, has an exponential distribution with parameter $\theta = \frac{1}{2}$ [*c.f.* § 9·8] or $Y = -2 \log X$, has chi-square distribution with $n = 2$ degrees of freedom [*c.f.* Chapter 15].

Example 9·24. *Show that for rectangular distribution : $f(x) = \frac{1}{2a}$, $-a < x < a$, m.g.f. about origin is $\frac{1}{at}(\sinh at)$. Also show that moments of even order are given by :*

$$\mu_{2n} = \frac{a^{2n}}{(2n+1)}$$

$= \frac{4b^2+4a^2+4ba-3b^2-8a^2-6ab}{12} \Rightarrow \frac{(a^2+b^2-ab)}{12} \Rightarrow \frac{(a-b)^2}{12}$

Solution. M.G.F. about origin is given by :

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx = \frac{1}{2a}\int_{-a}^{a} e^{tx}\,dx = \frac{1}{2at}\left|e^{tx}\right|_{-a}^{a} = \frac{1}{2at}(e^{at}-e^{-at}) = \frac{\sinh at}{at}$$

$$= \frac{1}{at}\left\{at + \frac{(at)^3}{3!} + \frac{(at)^5}{5!} + \ldots + \frac{(at)^{2n+1}}{(2n+1)!} + \ldots\right\} = 1 + \frac{a^2t^2}{3!} + \frac{a^4t^4}{5!} + \ldots + \frac{a^{2n}t^{2n}}{(2n+1)!} + \ldots$$

Since there are no terms with odd powers of t in $M_X(t)$, all moments of odd order about origin vanish, *i.e.*, μ'_{2n+1} (about origin) $= 0$.

In particular, μ_1' (about origin) $= 0$, *i.e.*, mean $= 0$

Thus μ_r' (about origin) $= \mu_r$ (Since mean is origin.)

Hence $\mu_{2n+1} = 0;\ n = 0, 1, 2, \ldots$, *i.e.*, all moments of odd order about mean vanish.

The moments of even order are given by :

$$\mu_{2n} = \text{coefficient of } \frac{t^{2n}}{(2n)!} \text{ in } M_X(t) = \frac{a^{2n}}{(2n+1)}$$

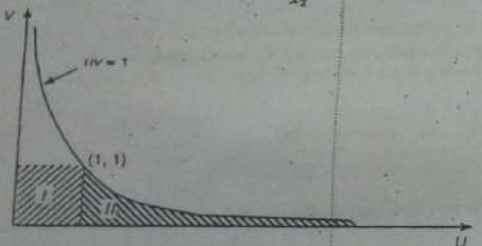
Example 9·25. *If X_1 and X_2 are independent rectangular variates on $[0, 1]$, find the distributions of : (i) X_1/X_2, (ii) X_1X_2, (iii) $X_1 + X_2$, and (iv) $X_1 - X_2$.*

Solution. We are given : $f_{X_1}(x_1) = f_{X_2}(x_2) = 1;\ 0 < x_1 < 1,\ 0 < x_2 < 1$.

Since X_1 and X_2 are independent, their joint *p.d.f.* is :

$$f(x_1, x_2) = f(x_1)\,f(x_2) = 1$$

(*i*) Let us transform to : $u = \frac{x_1}{x_2}$, $v = x_2$, *i.e.*, $x_1 = uv$, $x_2 = v$



$$J = \frac{\partial(x_1, x_2)}{\partial(u, v)} = \begin{vmatrix} v & 0 \\ u & 1 \end{vmatrix} = v$$

$x_1 = 0$ maps to $u = 0,\ v = 0$

$x_1 = 1$ maps to $uv = 1$ (Rectangular hyperbola)

$x_2 = 0$ maps to $v = 0$, and

$x_2 = 1$ maps to $v = 1$.

The joint *p.d.f.* of U and V becomes :

$$g(uv) = f(x_1, x_2)\,|J| = v;\ 0 < u < \infty,\ 0 < v < \infty$$

To obtain the marginal distribution of U, we have to integrate out v.

In region (*I*) : $g_1(u) = \int_0^1 v\,dv = \left|\frac{v^2}{2}\right|_0^1 = \frac{1}{2},\ 0 \le u \le 1$

In region (*II*) : $g_1(u) = \int_0^{1/u} v\,dv = \left|\frac{v^2}{2}\right|_0^{1/u} = \frac{1}{2u^2},\ 1 < u < \infty$

Hence the distribution $U = \frac{X_1}{X_2}$ is given by : $g(u) = \begin{cases} \frac{1}{2}, & 0 \le u \le 1 \\ \frac{1}{2u^2}, & 1 < u < \infty \end{cases}$

(*ii*) Let $u = x_1x_2,\ v = x_1 \Rightarrow x_1 = v,\ x_2 = \frac{u}{v}$ and $J = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}$,

$x_1 = 0$ maps to $v = 0$, $x_1 = 1$ maps to $v = 1$

$x_2 = 0$ maps to $u = 0$, and $x_2 = 1$ maps to $u = v$

Moreover, $v = \frac{u}{x_2} \Rightarrow v \ge u$ (Since $0 < x_2 < 1$).

The joint *p.d.f.* of U and V is : $g(u, v) = f(x_1, x_2)\,|J| = \frac{1}{v};\ 0 < u \le v < 1$

$$\therefore \quad g(u) = \int_u^1 g(u, v)\,dv = \int_u^1 \frac{1}{v}\,dv = \left|\log v\right|_u^1 = -\log u,\ 0 < u < 1$$

(*iii*) and (*iv*). Let $u = x_1 + x_2,\ v = x_1 - x_2$

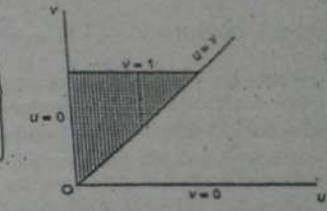
$$x_1 = \frac{u+v}{2}, \qquad x_2 = \frac{u-v}{2}$$

$x_1 = 0 \Rightarrow u + v = 0$, *i.e.*, $v = -u$

$x_2 = 0 \Rightarrow u - v = 0$, *i.e.*, $v = u$

$x_1 = 1 \Rightarrow u + v = 2$

$x_2 = 1 \Rightarrow u - v = 2$



and $J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$

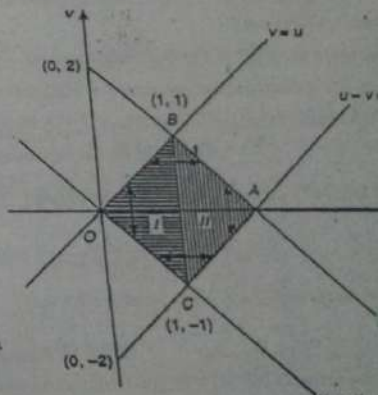
$$\therefore \quad g(u, v) = f(x_1, x_2)\,|J| = \tfrac{1}{2};\ 0 < u < 2,\ -1 < v < 1$$

In region (*I*)

$$g_1(u) = \int_{-u}^{u} \tfrac{1}{2}\,dv = \tfrac{1}{2}\left|v\right|_{-u}^{u} = u$$

and in region (*II*),

$$g_2(u) = \int_{u-2}^{2-u} \tfrac{1}{2}\,dv = \tfrac{1}{2}\left|v\right|_{u-2}^{2-u} = 2 - u$$



∴ Distribution $U = X_1 + X_2$, is given by :

$$g(u) = \begin{cases} u, & 0 < u < 1 \\ 2 - u, & 1 < u < 2 \end{cases}$$

For the distribution of V, we split the region as : OAB and OAC

In region OAB : $h_1(v) = \int_{v}^{2-v} \frac{1}{2}\,du = \frac{1}{2}(2 - v - v) = 1 - v,\ 0 < v < 1$

In region OAC : $h_2(v) = \int_{-v}^{2+v} \frac{1}{2}\,du = \frac{1}{2}[2(1 + v)] = 1 + v,\ -1 < v < 0$

Hence the distribution of $V = X_1 - X_2$ is given by : $h(v) = \begin{cases} 1 - v, & 0 < v < 1 \\ 1 + v, & -1 < v < 0 \end{cases}$

Example 9·26. *If X is a random variable with a continuous distribution function F, then prove that $F(X)$ has a uniform distribution on $[0, 1]$.*

Hence prove that : $P[a \le F(x) \le b] = b - a,\ 0 \le (a, b) \le 1$.

Solution. Since F is a distribution function, it is non-decreasing. Let $Y = F(X)$. Then the distribution function G of Y is given by :

$$G_Y(y) = P(Y \le y) = P[F(X) \le y] = P[X \le F^{-1}(y)],$$

the inverse exists, since F is non-decreasing and given to be continuous.

$\therefore$ $G_Y(y) = F[F^{-1}(y)]$, since F is the distribution function of X.

Thus $G_Y(y) = y$

Therefore the *p.d.f. of* $Y = F(X)$ is given by : $g_Y(y) = \frac{d}{dy}[G_Y(y)] = 1$

Since F is a distribution function $Y = F(X)$ takes the values in the range $[0, 1]$.

Hence $g_Y(y) = 1,\ 0 \le y \le 1 \quad \Rightarrow \quad Y$ is a uniform variate on $[0, 1]$.

Since $Y = F(X) \sim U[0, 1]$.

$$P[a \le F(X) \le b] = P[a \le Y \le b] = \int_a^b g(y)\,dy$$

$$= \int_a^b 1 \cdot dy = \left| y \right|_a^b = b - a.$$

Remark. Suppose X is a random variable with *p.d.f.*,

$$f_X(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{then } F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-x}, & \text{if } x \ge 0 \end{cases}$$

Then by above result $F(X) = 1 - e^{-X}$ is uniformly distributed on $[0, 1]$.

Example 9·27. *If X and Y are independent rectangular variates for the range $-a$ to a each, then show that the sum $X + Y = U$, has the probability density*

$$\varphi(u) = \begin{cases} \frac{2a + u}{4a^2}, & -2a \le u \le 0 \\ \frac{2a - u}{4a^2}, & 0 \le u \le 2a \end{cases}$$

Solution. Since X and Y are independent rectangular variates, each in the interval $(-a, a)$, we have

$$f_1(x) = \begin{cases} \frac{1}{2a}, & -a < x < a \\ 0, & \text{elsewhere} \end{cases} \quad \text{and} \quad f_2(y) = \begin{cases} \frac{1}{2a}, & -a < y < a \\ 0, & \text{elsewhere} \end{cases}$$

Hence by compound probability theorem, the joint probability differential of X and Y is given by :

$$dP(x, y) = f_1(x)\,f_2(y)\,dx\,dy = \frac{1}{4a^2}\,dx\,dy,\ -a < (x, y) < a$$

Let us define new variables U and V as follows :

$$u = x + y,\ v = x - y \quad \Rightarrow \quad x = \frac{u + v}{2} \quad \text{and} \quad y = \frac{u - v}{2}$$

Jacobian of the transformation J is given by :

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

Thus the probability differential of U and V becomes :

$$dG(u, v) = \frac{1}{4a^2} \cdot |J|\,du\,dv = \frac{1}{8a^2}\,du\,dv \qquad \ldots (*)$$

Integrating *w.r.* to v over specified range, we can find the distribution of U.

Let us consider the region to the left of v – axis, *i.e.*, to the left of the line AC. In this region, the values of v are bounded by the lines $x = -a$ and $y = -a$.

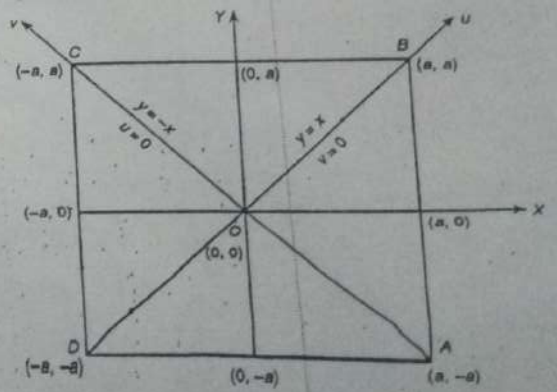
For fixed values of u,

$$x = -a \quad \Rightarrow \quad \frac{u + v}{2} = -a \quad \Rightarrow \quad v = -(u + 2a)$$

and $$y = -a \quad \Rightarrow \quad \frac{u - v}{2} = -a \quad \Rightarrow \quad v = (u + 2a)$$

Thus integrating (*) *w.r.* to v between the limits $-(u + 2a)$ and $(u + 2a)$, the distribution of U becomes :

$$\phi(u)\,du = \int_{-(u+2a)}^{u+2a} \frac{1}{8a^2}\,du\,dv = \frac{1}{8a^2}\left| v \right|_{-(u+2a)}^{u+2a} du = \frac{u + 2a}{4a^2}\,du$$



In the region to the left of v-axis, *i.e.*, below the line AC, u varies from the points $(x = -a, y = -a)$ to the point $(x = 0, y = 0)$ and since $u = x + y$, in this region u lies between $(-a - a)$ and $(0 + 0)$, *i.e.*, between $-2a$ to 0.

$$\therefore \quad \Phi(u) = \frac{u + 2a}{4a^2}, -2a \le u \le 0 \qquad \ldots (**)$$

In the region to the right of v-axis, *i.e.*, above the line AC, the values of v are bounded by the lines $x = a$ and $y = a$ and for fixed values of u,

$$x = a \quad \Rightarrow \quad \frac{u+v}{2} = a \quad \Rightarrow \quad v = 2a - u$$

$$y = a \quad \Rightarrow \quad \frac{u-v}{2} = a \quad \Rightarrow \quad v = -(2a - u)$$

In this region u varies from the point $(x = 0, y = 0)$ to the point $(x = a, y = a)$, *i.e.*, $u = x + y$ varies from 0 to $2a$. Thus integrating $(*)$ *w.r.* to. v between the limits $-(2a - u)$ to $(2a - u)$, we get the distribution of U as :

$$\phi(u)\,du = \int_{-(2a-u)}^{2a-u} \frac{1}{8a^2}\,du\,dv = \frac{1}{8a^2}|v|_{-(2a-u)}^{2a-u}\,du = \frac{2a-u}{4a^2}du,\ 0 \le u \le 2a \qquad \ldots (***)$$

Combining the results in $(**)$ and $(***)$, we get the required result.

(For an alternative and simpler solution, see Remark 5 to § 9·4, Triangular Distribution).

Example 9·28. *On the x-axis $(n + 1)$ points are taken independently between the origin and $x = 1$, all positions being equally likely. Show that probability that the $(k + 1)$ th of these points, counted from origin, lies in the interval $x - \frac{1}{2}dx$ to $x + \frac{1}{2}dx$ is :*

$$\binom{n}{k}(n+1)\,x^k(1-x)^{n-k}\,dx$$

Verify that integral of this expression from $x = 0$ to $x = 1$ is unity.

Solution. Here X is given to be a *r.v.* uniformly distributed on $[0,1]$.

$$\therefore \quad f_X(x) = 1,\ 0 \le x \le 1$$

$$\text{Now} \qquad P(0 < X < x) = \int_0^x f(x)\,dx = \int_0^x 1.\,dx = x \qquad \ldots (1)$$

$$\therefore \qquad P(X > x) = 1 - P(X \le x) = 1 - x \qquad \text{[From (1)]} \qquad \ldots (2)$$

$$\text{Also} \qquad P\left(x - \frac{dx}{2} < X < x + \frac{dx}{2}\right) = \int_{x-\frac{dx}{2}}^{x+\frac{dx}{2}} f(x)\,dx = dx$$

Required probability 'p' is given by : ... (3)

$p = P$ [out of $(n + 1)$ points, k points lie in closed interval $\left[0, x - \frac{dx}{2}\right]$ and out of the remaining $(n + 1 - k)$ points, $(n - k)$ points lie in $\left(x + \frac{dx}{2}, 1\right)$ and one point lies in $\left(x - \frac{dx}{2}, x + \frac{dx}{2}\right)$]

$$= \left[\binom{n+1}{k}x^k\right] \times \left[\binom{n+1-k}{n-k}(1-x)^{n-k}\right]dx\,,$$ on using (1), (2) and (3) respectively.

$$\therefore \quad p = \frac{(n+1)!}{k!\,(n+1-k)!} \cdot x^k \cdot \frac{(n+1-k)!}{(n-k)!} \cdot (1-x)^{n-k}\,dx = \binom{n}{k}(n+1)\,x^k(1-x)^{n-k}\,dx$$

Verification. $I = \int_0^1 \binom{n}{k}(n+1)\,x^k(1-x)^{n-k}\,dx = \binom{n}{k}(n+1)\,\beta(k+1, n-k+1)$

on using beta-Integral :

$$\int_0^1 x^{m-1}(1-x)^{n-1}\,dx = B(m,n) = \frac{\Gamma m\ \Gamma n}{\Gamma(m+n)};\ m > 0,\ n > 0.$$

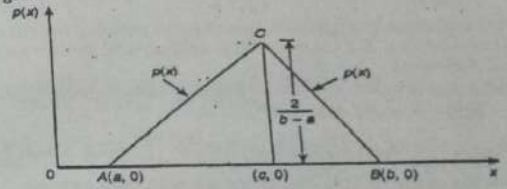
$$\therefore \qquad I = \frac{n!\,(n+1)}{k!\,(n-k)!} \cdot \frac{k!\,(n-k)!}{(n+1)!} = 1.$$

9·4. TRIANGULAR DISTRIBUTION

Definition. *A random variable X is said to have a triangular distribution in the interval (a, b), if its p.d.f. is given by :*

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}, & a < x \le c \\ \dfrac{2(b-x)}{(b-a)(b-c)}, & c < x < b \end{cases} \qquad \ldots (9\cdot21)$$

Remarks 1. We write $X \sim$ Trg. (a, b), with peak at $x = c$. The graph of the *p.d.f.* is shown in the following diagram :



2. The distribution is so called because the graph of its *p.d.f.* is a triangle with peak at $x = c$.

3. The *m.g.f.* of Trg (a, b) variate, with peak at $x = c$ is given by :

$$M_X(t) = \int_a^b e^{tx} f(x)\,dx = \left(\int_a^c + \int_c^b\right) e^{tx} f(x)\,dx$$

$$= \frac{2}{(b-a)(c-a)}\int_a^c e^{tx}(x-a)\,dx + \frac{2}{(b-a)(b-c)}\int_c^b e^{tx}(b-x)\,dx$$

$$= \frac{2}{t^2}\left\{\frac{e^{at}}{(a-b)(a-c)} + \frac{e^{ct}}{(c-a)(c-b)} + \frac{e^{bt}}{(b-a)(b-c)}\right\};\ a < b < c \qquad \ldots (9\cdot21a)$$

(On integration by parts)

4. In particular, taking $a = 0$, $c = 1$ and $b = 2$, in (9·21), the *p.d.f.* of the Trg $(0, 2)$ variate with peak at $x = 1$ is given by :

$$f(x) = \begin{cases} x; & 0 \le x \le 1 \\ 2 - x; & 1 \le x \le 2 \\ 0, & \text{otherwise} \end{cases} \qquad \ldots (9\cdot21$$

and its m.g.f. is : $M_X(t) = (e^t - 1)^3 / t^3$, ... (9·21c)
which is left as an exercise to the reader.

5. In particular, replacing a by $-2a$, b by $2a$ and c by 0, the *p.d.f.* of triangular distribution on the interval $(-2a, 2a)$ with peak at $x = 0$ is given by :

$$f(x) = \begin{cases} (2a + x)/4a^2 ; & -2a < x < 0 \\ (2a - x)/4a^2 ; & 0 < x < 2a \end{cases} \quad \ldots (9\cdot21d)$$

The *m.g.f.* of (9·21*d*) is given by :

$$M_X(t) = \int_{-2a}^{2a} e^{tx} f(x)\,dx = \frac{1}{4a^2}\left\{\int_{-2a}^{0} e^{tx}(2a + x)\,dx + \int_{0}^{2a} e^{tx}(2a - x)\,dx\right\}$$

$$= \frac{1}{4a^2}\left[e^{tx}\left\{\frac{2a + x}{t} - \frac{1}{t^2}\right\}\right]_{-2a}^{0} + \frac{1}{4a^2}\left[e^{tx}\left\{\frac{2a - x}{t} + \frac{1}{t^2}\right\}\right]_{0}^{2a} \quad \text{[On integrating by parts]}$$

$$= \frac{1}{4a^2}\left[-\frac{2}{t^2} + \frac{1}{t^2}\left\{e^{2at} + e^{-2at}\right\}\right] = \frac{1}{4a^2t^2}\left\{e^{2at} + e^{-2at} - 2\right\} = \left\{\frac{1}{2at}(e^{at} - e^{-at})\right\}^2$$

Aliter. We may obtain (9·21) directly from (9·2*a*) on replacing a by $-2a$, b by $2a$ and c by 0.

Example 9·29. *If X and Y are i.i.d. U [−a, a] variates, find the p.d.f. of Z = X + Y and identify the distribution.*

Solution. Since X and Y are *i.i.d.* U $[-a, a]$, we have : [*c.f.* § 9·3·2],

$$M_X(t) = M_Y(t) = (e^{at} - e^{-at})/(2at) \quad \ldots (*)$$

$$M_{X+Y}(t) = M_X(t)\,M_Y(t) = \left\{\frac{1}{2at}(e^{at} - e^{-at})\right\}^2, \text{ since } X \text{ and } Y \text{ are independent.}$$

But, this is the *m.g.f.* of Trg $(-2a, 2a)$ variate with peak at $x = 0$.
[*c.f.* Remark 5, equation (9·21*e*)]

Hence by uniqueness theorem of *m.g.f.*, $Z = X + Y \sim$ Trg $(-2a, 2a)$ with *p.d.f.* as given in (9·21*d*), Remark 5.

Aliter. $M_{X+Y}(t) = \frac{1}{4a^2t^2}(e^{2at} - 2 + e^{-2at})$ [From (**)]

$$= \frac{2}{t^2}\left[\frac{e^{-2at}}{(-2a - 0)(-2a - 2a)} + \frac{e^{0\cdot t}}{(0 + 2a)(0 - 2a)} + \frac{e^{2at}}{(2a - 0)(2a + 2a)}\right]$$

which is of the form (9·21*a*), [*c.f.* Remark 3], with a replaced by $-2a$ and b replaced by $2a$ and c by 0. Hence $X + Y \sim$ Trg $(-2a, 2a)$ with peak at $x = 0$ and *p.d.f.* $p(x)$ given in (9·21*d*).

Remarks 1. The distribution of $X + Y$ has also been obtained in *Example* 9·27.

2. Similarly we can find the distribution of $X - Y$.

$$M_{X-Y}(t) = M_X(t).\,M_Y(-t) = \left\{\frac{1}{2at}(e^{at} - e^{-at})\right\}^2 \quad \text{[From (*)]}$$

$X - Y \sim$ Trg $(-2a, 2a)$, with peak at $x = 0$.

9·5. GAMMA DISTRIBUTION

Definition. *A r.v. X is said to have a gamma distribution with parameter $\lambda > 0$, if its p.d.f. is given by :*

$$f(x) = \begin{cases} \dfrac{e^{-x}x^{\lambda-1}}{\Gamma(\lambda)} ; & \lambda > 0,\ 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \quad \ldots (9\cdot22)$$

Remarks 1. X is known as a Gamma variate with parameter λ and referred to as a $\gamma(\lambda)$ variate.

2. The function $f(x)$ defined above represents a probability function, since

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

3. A continuous random variable X having the following *p.d.f.* is said to have a gamma distribution with two parameters a and λ.

$$f(x) = \begin{cases} \dfrac{a^\lambda}{\Gamma(\lambda)}e^{-ax}x^{\lambda-1} ; & a > 0, \lambda > 0,\ 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \quad \ldots (9\cdot22a)$$

Here $X \sim \gamma(a, \lambda)$. Taking $a = 1$ in (9·22*a*), we get (9·22). Hence we may write $X \sim \gamma(\lambda) = \gamma(1, \lambda)$.

4. The *cumulative distribution function*, called incomplete gamma function is defined as :

$$F_X(x) = \begin{cases} \int_0^x f(u)\,du = \dfrac{1}{\Gamma(\lambda)}\int_0^x e^{-u}u^{\lambda-1}\,du, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad \ldots (9\cdot22b)$$

9·5·1. M.G.F. of Gamma Distribution. M.G.F. about origin is given by :

$$M_x(t) = E(e^{tX}) = \int_0^\infty e^{tx}f(x)\,dx = \frac{1}{\Gamma(\lambda)}\int_0^\infty e^{tx}e^{-x}x^{\lambda-1}\,dx$$

$$= \frac{1}{\Gamma(\lambda)}\int_0^\infty e^{-(1-t)x}x^{\lambda-1}\,dx = \frac{1}{\Gamma(\lambda)}\cdot\frac{\Gamma(\lambda)}{(1-t)^\lambda}, \quad |t| < 1$$

$$\therefore \quad M_X(t) = (1 - t)^{-\lambda}, \quad |t| < 1 \quad \ldots (9\cdot23)$$

9·5·2. Cumulant Generating Function of Gamma Distribution. The cumulant generating function $K_X(t)$ is given by :

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

$$= \lambda\left(t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \ldots\right)$$

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

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

Hence if $X \sim \gamma(\lambda)$, Mean = Variance = λ.

$\mu_3 = \kappa_3 =$ Coefficient of $\frac{t^3}{3!}$ in $K_X(t) = 2\lambda$

$\kappa_4 =$ Coefficient of $\frac{t^4}{4!}$ in $K_X(t) = 6\lambda \implies \mu_4 = \kappa_4 + 3\kappa_2^2 = 6\lambda + 3\lambda^2$

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{4\lambda^2}{\lambda^3} = \dfrac{4}{\lambda}$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3 + \dfrac{6}{\lambda}$.

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

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

$$M_Z(t) = \exp(-\mu t/\sigma) M_X(t/\sigma) = \exp(-\mu t/\sigma) \left(1 - \frac{t}{\sigma}\right)^{-\lambda} = e^{-t\sqrt{\lambda}} \left(1 - \frac{t}{\sqrt{\lambda}}\right)^{-\lambda}$$

$$\Rightarrow \quad K_Z(t) = \sqrt{\lambda}\, t - \lambda \log\left(1 - \frac{t}{\sqrt{\lambda}}\right) = -\sqrt{\lambda}\, t - \lambda \left(\frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{3\lambda^{3/2}} + \ldots\right)$$

$$= -\sqrt{\lambda}\, t + \sqrt{\lambda}\, t + \frac{t^2}{2} + \phi(\lambda^{-1/2}),$$

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

$$\therefore \lim_{\lambda \to \infty} K_Z(t) = \frac{t^2}{2} \quad \Rightarrow \quad \lim_{\lambda \to \infty} M_Z(t) = \exp(t^2/2)$$

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

3. For the two parameter gamma distribution (9·22 *a*), we have

$$M_X(t) = \left(1 - \frac{t}{a}\right)^{-\lambda}; t < a. \qquad \ldots(9\cdot23\,a)$$

Proof is left as an excercise to the reader.

Also $$K_X(t) = -\lambda \log\left(1 - \frac{t}{a}\right) = \lambda\left\{\frac{t}{a} + \frac{1}{2}\left(\frac{t}{a}\right)^2 + \frac{1}{3}\left(\frac{t}{a}\right)^3 + \ldots\right\}; t < a$$

$\therefore$ Mean $= \kappa_1 = \lambda/a$ and Variance $= \kappa_2 = \lambda/a^2 =$ Mean$/a$ $\qquad \ldots(9\cdot23\,b)$

Hence Variance > Mean if $a < 1$; Variance = Mean if $a = 1$; and Variance < Mean if $a > 1$.

9·5·3. Additive Property of Gamma Distribution. *The sum of independent Gamma variates is also a Gamma variate. More precisely, if $X_1, X_2, \ldots, X_k$ are independent Gamma variates with parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$ respectively then $X_1 + X_2 + \ldots + X_k$ is also a Gamma variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_k$.*

Proof. Since X_i is a $\gamma(\lambda_i)$ variate, $M_{X_i}(t) = (1-t)^{-\lambda_i}$

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

$$M_{X_1 + X_2 + \ldots X_k}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_k}(t), \qquad (\because X_1, X_2, \ldots, X_k \text{ are independent.})$$

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

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

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

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

9·6. BETA DISTRIBUTION OF FIRST KIND

Definition. A r.v. X is said to have a beta distribution of first kind with parameters μ and ν ($\mu > 0$, $\nu > 0$) if its p.d.f. is given by :

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

[where $B(\mu, \nu)$ is the Beta function]. $\qquad \ldots(9\cdot24)$

The r.v. X is known as a Beta variate of the first kind with parameters μ and ν and is referred to as $\beta_1(\mu, \nu)$ variate.

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

$$F(x) = \begin{cases} 0, x < 0 \\ \displaystyle\int_0^x \frac{1}{B(\mu, \nu)} u^{\mu-1} (1-u)^{\nu-1}\, du; 0 < x < 1, (\mu, \nu) > 0 \\ 1, x > 1 \end{cases} \qquad \ldots(9\cdot24a)$$

2. In particular, if we take $\mu = 1$ $\nu = 1$ in (9·24), we get

$$f(x) = \frac{1}{B(1,1)} = 1, 0 < x < 1 \qquad \ldots(9\cdot24b)$$

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

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

9·6·1. Constants of Beta Distribution of First Kind.

$$\mu_r' = \int_0^1 x^r f(x)\, dx = \frac{1}{B(\mu, \nu)} \int_0^1 x^{\mu+r-1} (1-x)^{\nu-1}\, dx = \frac{1}{B(\mu, \nu)} B(\mu + r, \nu)$$

$$= \frac{\Gamma(\mu + r)\, \Gamma(\nu)}{\Gamma(\mu + r + \nu)} \cdot \frac{\Gamma(\mu + \nu)}{\Gamma(\mu)\, \Gamma(\nu)} = \frac{\Gamma(\mu + r)\, \Gamma(\mu + \nu)}{\Gamma(\mu + r + \nu)\, \Gamma(\mu)} \qquad \ldots(9\cdot24c)$$

In particular

$$\text{Mean} = \mu_1' = \frac{\Gamma(\mu + 1)}{\Gamma(\mu + \nu + 1)} \cdot \frac{\Gamma(\mu + \nu)}{\Gamma(\mu)} = \frac{\mu\, \Gamma(\mu)\, \Gamma(\mu + \nu)}{(\mu + \nu)\, \Gamma(\mu + \nu)\, \Gamma(\mu)} = \frac{\mu}{\mu + \nu} \qquad \ldots(9\cdot24\,d)$$

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

$$\mu_2' = \frac{\Gamma(\mu + 2) \cdot \Gamma(\mu + \nu)}{\Gamma(\mu + \nu + 2)\, \Gamma(\mu)} = \frac{(\mu + 1)\, \mu\, \Gamma(\mu)\, \Gamma(\mu + \nu)}{(\mu + \nu + 1)(\mu + \nu)\, \Gamma(\mu + \nu)\, \Gamma(\mu)} = \frac{\mu(1 + \mu)}{(\mu + \nu)(\mu + \nu + 1)}$$

Hence $$\mu_2 = \mu_2' - \mu_1'^2 = \frac{\mu(1 + \mu)}{(\mu + \nu)(\mu + \nu + 1)} - \left(\frac{\mu}{\mu + \nu}\right)^2$$

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

Similarly, we have

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

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

$$= \frac{3\mu\nu\left[\mu\nu(\mu + \nu - 6) + 2(\mu + \nu)^2\right]}{(\mu + \nu)^4 (\mu + \nu + 1)(\mu + \nu + 2)(\mu + \nu + 3)}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4(\nu - \mu)^2 (\mu + \nu + 1)}{\mu\nu(\mu + \nu + 2)^2} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(\mu + \nu + 1)\mu\nu(\mu + \nu - 6) + 2(\mu + \nu)^2}{\mu\nu(\mu + \nu + 2)(\mu + \nu + 3)}$$

The harmonic mean H is given by :

$$\frac{1}{H} = \int_0^1 \frac{1}{x} f(x)\, dx = \frac{1}{B(\mu, \nu)} \int_0^1 x^{\mu-2} (1-x)^{\nu-1}\, dx$$

$$= \frac{1}{B(\mu, \nu)} B(\mu - 1, \nu) = \frac{\Gamma(\mu - 1)\, \Gamma(\nu)}{\Gamma(\mu + \nu - 1)} \cdot \frac{\Gamma(\mu + \nu)}{\Gamma(\mu)\, \Gamma(\nu)}$$

$$= \frac{\Gamma(\mu-1)\,(\mu+\nu-1)\,\Gamma(\mu+\nu-1)}{\Gamma(\mu+\nu-1)\,(\mu-1)\,\Gamma(\mu-1)} = \frac{\mu+\nu-1}{\mu-1}$$

$$\therefore \quad H = \frac{\mu-1}{\mu+\nu-1} \qquad \ldots(9\cdot24c)$$

9.7. BETA DISTRIBUTION OF SECOND KIND

Definition. *A r.v. X is said to have a beta distribution of the second kind with parameters μ and υ, $(\mu > 0, \upsilon > 0)$, if its p.d.f. is given by :*

$$f(x) = \begin{cases} \dfrac{1}{B(\mu,\nu)} \cdot \dfrac{x^{\mu-1}}{(1+x)^{\mu+\nu}}; (\mu,\nu) > 0,\ 0 < x < \infty \\ 0, \text{ otherwise} \end{cases} \qquad \ldots(9\cdot25)$$

Remarks 1. The r.v. X is known as a Beta variate of second kind with parameters μ and ν and is denoted as $\beta_2(\mu, \nu)$ variate.

2. Beta distribution of second kind is transformed to Beta distribution of first kind by the transformation : $\quad 1 + x = \dfrac{1}{y} \implies y = \dfrac{1}{1+x} \qquad \ldots(*)$

Thus, if $X \sim \beta_2(\mu, \nu)$, then Y defined in (*) is a $\beta_1(\mu, \nu)$.

The proof is left as an exercise to the reader.

9.7.1. Constants of Beta Distribution of Second kind.

$$\mu_r' = \int_0^\infty x^r f(x)\,dx = \frac{1}{B(\mu,\nu)} \int_0^\infty \frac{x^{\mu+r-1}}{(1+x)^{\mu+\nu}}\,dx$$

$$= \frac{1}{B(\mu,\nu)} \int_0^\infty \frac{x^{(\mu+r)-1}}{(1+x)^{\mu+r+\nu-r}}\,dx = \frac{1}{B(\mu,\nu)} \cdot B(\mu+r, \nu-r)$$

$$= \frac{\Gamma(\mu+r)\,\Gamma(\nu-r)}{\Gamma(\mu+\nu)} \cdot \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\,\Gamma(\nu)} = \frac{\Gamma(\mu+r)\,\Gamma(\nu-r)}{\Gamma(\mu)\,\Gamma(\nu)}; \upsilon > r$$

In particular

$$\mu_1' = \frac{\Gamma(\mu+1)\,\Gamma(\nu-1)}{\Gamma(\mu)\,\Gamma(\nu)} = \frac{\mu\,\Gamma(\mu)\,\Gamma(\nu-1)}{\Gamma(\mu)\,(\nu-1)\,\Gamma(\nu-1)} = \frac{\mu}{\nu-1}, \upsilon > 1$$

$$\mu_2' = \frac{\Gamma(\mu+2)\,\Gamma(\nu-2)}{\Gamma(\mu)\,\Gamma(\nu)} = \frac{(\mu+1)\,\mu\,\Gamma(\mu)\,\Gamma(\nu-2)}{\Gamma(\mu)\,(\nu-1)\,(\nu-2)\,\Gamma(\nu-2)} = \frac{\mu(\mu+1)}{(\nu-1)(\nu-2)}, \upsilon > 2$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \frac{\mu(\mu+1)}{(\nu-1)(\nu-2)} - \left(\frac{\mu}{\nu-1}\right)^2$$

$$= \frac{\mu}{\nu-1}\left[\frac{(\nu-1)(\mu+1) - \mu(\nu-2)}{(\nu-1)\,(\nu-2)}\right] = \frac{\mu(\mu+\nu-1)}{(\nu-1)^2(\nu-2)}$$

The harmonic mean H is given by :

$$\frac{1}{H} = E\left(\frac{1}{X}\right) = \int_0^\infty \frac{1}{x} \cdot f(x)\,dx = \frac{1}{B(\mu,\nu)} \int_0^\infty \frac{x^{\mu-2}}{(1+x)^{\mu+\nu}}\,dx$$

$$= \frac{1}{B(\mu,\nu)} \int_0^\infty \frac{x^{\mu-1+1}}{(1+x)^{\mu-1+\nu+1}}\,dx = \frac{1}{B(\mu,\nu)} \cdot B(\mu-1, \nu+1), \mu > 1$$

$$= \frac{\Gamma(\mu-1)\,\Gamma(\upsilon+1)}{\Gamma(\mu+\upsilon)} \cdot \frac{\Gamma(\mu+\upsilon)}{\Gamma(\mu)\,\Gamma(\upsilon)} = \frac{\Gamma(\mu-1)\,\upsilon\,\Gamma(\upsilon)}{(\mu-1)\,\Gamma(\mu-1)\,\Gamma(\upsilon)} = \frac{\upsilon}{\mu-1}$$

Hence $\quad H = \dfrac{\mu-1}{\nu}$

Example 9·30. *The daily consumption of milk in a city, in excess of 20,000 litres, is approximately distributed as a Gamma variate with parameters $a = \frac{1}{10,000}$ and $\lambda = 2$. The city has a daily stock of 30,000 litres. What is the probability that the stock is insufficient on a particular day ?*

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

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

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

$$p = P(X > 30{,}000) = P(Y > 10{,}000) = \int_{10{,}000}^\infty g(y)\,dy = \int_{10{,}000}^\infty \frac{y\, e^{-y/10{,}000}}{(10{,}000)^2}\,dy$$

$$= \int_1^\infty z\, e^{-z}\,dz \qquad [\text{Taking } z = y/10{,}000]$$

Integrating by parts, $p = \left| -z\, e^{-z} \right|_1^\infty + \int_1^\infty e^{-z}\,dz = e^{-1} - \left| e^{-z} \right|_1^\infty = e^{-1} + e^{-1} = \frac{2}{e}$

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

$\int_0^\infty \frac{e^{-x}\, x^{n-1}}{\Gamma n}\,dx$, which have been tabulated for different values of α and n.

Example 9·31. *If $X \sim N(\mu, \sigma^2)$, obtain the p.d.f. of : $U = \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$.*

Solution. Since $X \sim N(\mu, \sigma^2) \implies z = \dfrac{x-\mu}{\sigma} \sim N(0, 1)$ with p.d.f. :

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right); -\infty < z < \infty. \qquad \ldots(i)$$

The distribution function $G(\cdot)$ of $U = \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$ is given by :

$$G_U(u) = P(U \le u) = P\left[\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 \le u\right]$$

$$= P(Z^2 \le 2u) = P\left(-\sqrt{2u} \le Z \le \sqrt{2u}\right), \quad \text{where } Z \sim N(0, 1)$$

$$= P(Z \le \sqrt{2u}) - P(Z \le -\sqrt{2u})$$
$$= \Phi(\sqrt{2u}) - \Phi(-\sqrt{2u}) \qquad \ldots (ii)$$

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

Differentiating w.r. to u, the *p.d.f.* $g(\cdot)$ of U is given by :

$$g(u) = \phi(\sqrt{2u}) \cdot \frac{d}{du}(\sqrt{2u}) - \phi(-\sqrt{2u}) \cdot \frac{d}{du}(-\sqrt{2u})$$

[$\phi(\cdot)$ is p.d.f. of SNV Z]

$$= \frac{1}{\sqrt{2u}} \left[\phi(\sqrt{2u}) + \phi(-\sqrt{2u})\right]$$
$$= \frac{1}{\sqrt{2u}} \cdot 2\,\phi(\sqrt{2u}) \qquad [\because \phi(u) \text{ is even function of } u]$$
$$= \sqrt{\frac{2}{u}} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot 2u\right) \qquad \text{[From (i)]}$$
$$= \frac{1}{\Gamma(1/2)} e^{-u} \cdot u^{(1/2)-1}, \; u \ge 0$$

$$\left[\because \sqrt{\pi} = \Gamma(1/2) \text{ and } u = \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \ge 0\right]$$

which is the *p.d.f.* of gamma distribution with parameter $\frac{1}{2}$.

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

Example 9-32. *Show that the mean value of positive square root of a $\gamma(\mu)$ variate is $\Gamma(\mu + \frac{1}{2})/\Gamma(\mu)$. Hence prove that the mean deviation of a normal variate from its mean is $\sqrt{2/\pi}$ σ, where σ is the standard deviation of the distribution.*

Solution. Let X be a $\gamma(\mu)$ variate. Then $f(x) = \frac{e^{-x} x^{\mu-1}}{\Gamma(\mu)}$; $\mu > 0, 0 < x < \infty$

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

If $X \sim N(\mu, \sigma^2)$, then $U = \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$ is a $\gamma\left(\frac{1}{2}\right)$ variate. (*c.f.* Example 9-31)

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

Hence mean deviation of X about mean is given by :

$$E|X-\mu| = E(\sqrt{2}\,\sigma\,U^{1/2}) = \sqrt{2}\,\sigma\,E(U^{1/2})$$
$$= \sqrt{2}\,\sigma\,\frac{\Gamma\left(\frac{1}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} = \frac{\sqrt{2}\,\sigma}{\sqrt{\pi}} = \sigma\sqrt{2/\pi}. \qquad \text{[Using (*) with } \mu = \tfrac{1}{2}\text{]}$$

Example 9-33. *If X and Y are independent Gamma variates with parameters μ and ν respectively, show that the variables $U = X + Y$, $Z = \frac{X}{X+Y}$ are independent and that U is a $\gamma(\mu + \nu)$ variate and Z is a $\beta_1(\mu, \nu)$ variate.*

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

$$f_1(x)\,dx = \frac{1}{\Gamma(\mu)} e^{-x} x^{\mu-1}\,dx\,; 0 < x < \infty, \mu > 0$$
$$f_2(y)\,dy = \frac{1}{\Gamma(\nu)} e^{-y} y^{\nu-1}\,dy\,; 0 < y < \infty, \nu > 0$$

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

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

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

Jacobian of transformation J is given by :

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

As X and Y range from 0 to ∞, u ranges from 0 to ∞ and z from 0 to 1 $\left(\because \frac{x}{x+y} \le 1\right)$.

Hence the joint distribution of U and Z is given by :

$$dG(u, z) = g(u, z)\,du\,dz = \frac{1}{\Gamma(\mu)\,\Gamma(\nu)} e^{-u} (uz)^{\mu-1} [u(1-z)]^{\nu-1} |J|\,du\,dz$$
$$= \frac{1}{\Gamma(\mu)\,\Gamma(\nu)} \cdot e^{-u} u^{\mu+\nu-1} z^{\mu-1} (1-z)^{\nu-1}\,du\,dz$$
$$= \left\{\frac{e^{-u} u^{\mu+\nu-1}}{\Gamma(\mu+\nu)}\,du\right\}\left\{\frac{1}{B(\mu, \nu)} z^{\mu-1} (1-z)^{\nu-1}\,dz\right\}$$
$$= [g_1(u)\,du]\,[g_2(z)\,dz], \text{ (say)}, \qquad \ldots (*)$$

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

and $g_2(z) = \frac{1}{B(\mu, \nu)} z^{\mu-1} (1-z)^{\nu-1}$, $0 < z < 1$ $\qquad \ldots (**)$

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

Example 9-34. *If X and Y are independent Gamma variates with parameters μ and ν respectively, show that $U = X + Y$, $Z = \frac{X}{Y}$ are independent and that U is a $\gamma(\mu + \nu)$ variate and Z is a $\beta_2(\mu, \nu)$ variate.*

Solution. As in example 9-33, we have

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

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

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

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

$$dG(u,z) = \frac{1}{\Gamma(\mu)\Gamma(\nu)} e^{-\left(\frac{uz}{1+z}\right)} \left(\frac{uz}{1+z}\right)^{\mu-1} \left(\frac{u}{1+z}\right)^{\nu-1} |J| \, du\,dz$$

$$= \left[\frac{e^{-u} u^{\mu+\nu-1}}{\Gamma(\mu+\nu)} du\right]\left[\frac{1}{B(\mu,\nu)} \cdot \frac{z^{\mu-1}}{(1+z)^{\mu+\nu}} dz\right]; 0 < u < \infty, 0 < z < \infty$$

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

Remark. The above two examples lead to the following important results.

If X is a $\gamma(\mu)$ variate and Y is an independent $\gamma(\nu)$ variate, then

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

(ii) $\frac{X}{Y}$ is a $\beta_2(\mu, \nu)$ variate, *i.e.*, the ratio of two independent Gamma variates is β_2 – variate.

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

Example 9-35. *Let X and Y have joint p.d.f. :*

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

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

Solution. Let $u = \frac{x}{x+y}$ and $v = x + y \Rightarrow x = uv, y = v - x = v - uv = v(1-u)$.

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

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

$$p(u,v) = g(x,y) \cdot |J| = \frac{1}{\Gamma 4 \Gamma 5} e^{-v} \cdot (uv)^3 [v(1-u)]^4 \times v$$

$$= \frac{1}{\Gamma 4 \Gamma 5} e^{-v} \cdot v^8 \cdot u^3 (1-u)^4; \; 0 \le u \le 1, v > 0$$

($\because u = \frac{x}{x+y} < 1$ and since $x > 0, y > 0$, we have $0 < u < 1$ and $v = x + y \ge 0$)

$$= \left(\frac{1}{\Gamma 9} e^{-v} v^8\right)\left\{\frac{\Gamma 9}{\Gamma 4 \Gamma 5} u^3 (1-u)^4\right\}; \; 0 < u < 1, v > 0$$

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

where $p_1(v) = \frac{1}{\Gamma 9} e^{-v} v^8; v > 0$ and $p_2(u) = \frac{\Gamma 9}{\Gamma 4 \Gamma 5} u^3 (1-u)^4; 0 < u < 1 \qquad \ldots(**)$

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

$U = \frac{X}{X+Y} \sim \beta_1(4, 5)$. *i.e.*, U is a Beta variate of first kind with parameters (4, 5).

Alter. We have

$$g(x,y) = \frac{1}{\Gamma 4 \Gamma 5} e^{-(x+y)} x^3 y^4 = \left[\frac{1}{\Gamma 4} e^{-x} x^3\right]\left[\frac{1}{\Gamma 5} e^{-y} y^4\right]$$

$$= g_1(x)\, g_2(y); x > 0, y > 0$$

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

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

$$\therefore \quad E(U) = \int_0^1 u \cdot p_2(u)\, du = \frac{1}{B(4,5)} \cdot \int_0^1 u^4 (1-u)^4 \, du$$

$$= \frac{1}{B(4,5)} \cdot B(5,5) \qquad \text{[Using Beta integral]}$$

$$= \frac{\Gamma 9}{\Gamma 4 \Gamma 5} \times \frac{\Gamma 5 \Gamma 5}{\Gamma 10} = \frac{\Gamma 9 \cdot 4\Gamma 4}{\Gamma 4 \cdot 9\Gamma 9} = \frac{4}{9}$$

$$E(U^2) = \frac{1}{B(4,5)} \int_0^1 u^2 \cdot u^3 (1-u)^4 \, du$$

$$= \frac{1}{B(4,5)} \times B(6,5) = \frac{\Gamma 9}{\Gamma 4 \Gamma 5} \times \frac{\Gamma 6 \Gamma 5}{\Gamma 11} = \frac{5 \times 4}{10 \times 9} = \frac{2}{9}$$

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

Example 9-36. *A random sample of size n is taken from a population with distribution*

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

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

Solution.

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

$$= \frac{1}{\Gamma(\lambda) a^\lambda} \int_0^\infty \exp\left\{-\left(\frac{1}{a} - t\right) x\right\} x^{\lambda-1} \, dx = \frac{1}{\Gamma(\lambda) a^\lambda} \cdot \frac{\Gamma(\lambda)}{\left(\frac{1}{a} - t\right)^\lambda} = (1 - at)^{-\lambda}$$

$$\therefore \quad M_{\bar{X}}(t) = M_{(X_1 + X_2 + \ldots + X_n)/n}(t) = M_{X_1 + X_2 + \ldots + X_n}(t/n) \qquad [\because M_{cX}(t) = M$$

$$= M_{X_1}(t/n)\, M_{X_2}(t/n) \ldots M_{X_n}(t/n). \qquad \text{(Since } X_1, X_2, \ldots, X_n \text{ are indepe}$$

Hence on using (*), we get $M_{\bar{X}}(t) = \left[\left(1 - \frac{at}{n}\right)^{-\lambda}\right]^n = \left(1 - \frac{t}{n/a}\right)^{-n\lambda}$

which is the *m.g.f.* of a Gamma distribution (*c.f.* Remark 3, § 9·5·2). Hen uniqueness theorem of *m.g.f.*, $\bar{X} \sim \gamma(n/a, n\lambda)$ with *p.d.f.* :

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

Example 9·37. A sample of n values is drawn from a population whose probability density is ae^{-ax}, $(x \geq 0, a > 0)$. If $\overline{X}$ is mean of the sample, show that $na\overline{X}$ is a $\gamma(n)$ variate and prove that $E(\overline{X}) = \frac{1}{a}$ and S.E. of $(\overline{X}) = \frac{1}{a\sqrt{n}}$.

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

$$M_X(t) = \int_0^\infty e^{tx} f(x)\,dx = a\int_0^\infty e^{-(a-t)x}\,dx = a\left|\frac{e^{-(a-t)x}}{-(a-t)}\right|_0^\infty = \frac{a}{a-t}, \quad (a > t) \quad \ldots (*)$$

$$an\overline{X} = an\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = a(X_1 + X_2 + \ldots + X_n)$$

$$M_{an\overline{X}}(t) = M_a(X_1 + X_2 + \ldots + X_n)(t)$$
$$= M_{X_1}(at) \cdot M_{X_2}(at) \ldots M_{X_n}(at),$$
(Since the sample values are independent.)

$$\therefore M_{an\overline{X}}(t) = \prod_{i=1}^{n} M_{X_i}(at) = [M_{X_i}(at)]^n, \text{ (Since } X_1, X_2, \ldots, X_n \text{ are identically distributed.)}$$

$$\therefore M_{an\overline{X}}(t) = \left(\frac{1}{1-t}\right)^n = (1-t)^{-n}, \qquad \text{[From (*)]}$$

which is the *m.g.f.* of a $\gamma(n)$ variate.

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

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

$$E(an\overline{X}) = n \quad \Rightarrow \quad an\,E(\overline{X}) = n \quad \Rightarrow \quad E(\overline{X}) = \frac{1}{a}$$

and
$$V(an\overline{X}) = n \quad \Rightarrow \quad a^2 n^2 V(\overline{X}) = n \quad \Rightarrow \quad V(\overline{X}) = \frac{1}{na^2}$$

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

Note. For concept of Standard Error (S.E.) see Chapter Fourteen.

***Example 9·38.** Let $X \sim \beta_1(\mu, \upsilon)$ and $Y \sim \gamma(\lambda, \mu + \upsilon)$ be independent random variables, $(\mu, \upsilon, \lambda > 0)$. Find a p.d.f. for XY and identify its distribution.*

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

$$f(x, y) = \frac{1}{B(\mu, \upsilon)} \cdot x^{\mu-1}(1-x)^{\upsilon-1} \times \frac{\lambda^{\mu+\upsilon}}{\Gamma(\mu+\upsilon)} e^{-\lambda y} y^{\mu+\upsilon-1}; 0 < x < 1, 0 < y < \infty$$

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

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

Jacobian of transformation J is given by : $J = \frac{\partial(x, y)}{\partial(u, z)} = \begin{vmatrix} 0 & 1 \\ \frac{1}{z} & -\frac{u}{z^2} \end{vmatrix} = -\frac{1}{z}$

Thus the *p.d.f.* of U and Z becomes :

$$g(u, z) = \frac{\lambda^{\mu+\upsilon}}{B(\mu, \upsilon)\Gamma(\mu+\upsilon)} \cdot (z)^{\mu-1}(1-z)^{\upsilon-1} e^{-\lambda u/z} \left(\frac{u}{z}\right)^{\mu+\upsilon-1} |J|\,; 0 < u < \infty, 0 < z < 1$$

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

$$g_1(u) = \frac{\lambda^{\mu+\upsilon} u^{\mu+\upsilon-1}}{\Gamma(\mu)\Gamma(\upsilon)} \int_0^1 \frac{(1-z)^{\upsilon-1} e^{-\lambda u/z}}{z^{\upsilon+1}}\,dz$$

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

$$= \frac{\lambda^{\mu+\upsilon} u^{\mu+\upsilon-1}}{\Gamma(\mu)\Gamma(\upsilon)} \int_0^\infty t^{\upsilon-1} e^{-\lambda u(1+t)}\,dt, \qquad \left(\frac{1}{z} - 1 = t\right)$$

$$= \frac{\lambda^{\mu+\upsilon} u^{\mu+\upsilon-1} e^{-\lambda u}}{\Gamma(\mu)\Gamma(\upsilon)} \int_0^\infty e^{-\lambda u t} t^{\upsilon-1}\,dt = \frac{\lambda^{\mu+\upsilon} u^{\mu+\upsilon-1} e^{-\lambda u}}{\Gamma(\mu)\Gamma(\upsilon)} \cdot \frac{\Gamma(\upsilon)}{(\lambda u)^\upsilon}$$

$$= \frac{\lambda^\mu}{\Gamma(\mu)} \cdot e^{-\lambda u} u^{\mu-1}, 0 < u < \infty$$

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

***Example 9·39.** Let $p \sim \beta_1(a, b)$ where a and b are positive integers. After one observes p, one secures a coin for which the probability of head is p. This coin is flipped n times. Let X denote the number of heads which result. Find $P(X = k)$ for, $k = 0, 1, 2, \ldots, n$. Express the answer in terms of binomial coefficients.*

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

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

$P(X = k \mid$ the probability of success in a single trial is $p) = \binom{n}{k} p^k q^{n-k}, q = 1 - p$

$$P(X = k) = \int_0^1 f(p)\,P(X = k \mid p)\,dp$$

$$= \int_0^1 \frac{1}{B(a, b)} p^{a-1}(1-p)^{b-1} \cdot \binom{n}{k} p^k (1-p)^{n-k}\,dp$$

$$= \frac{\binom{n}{k}}{B(a, b)} \int_0^1 p^{a+k-1}(1-p)^{n+b-k-1}\,dp = \frac{\binom{n}{k} B(a+k, n+b-k)}{B(a, b)}$$

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

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

***Example 9·40.** Given the Incomplete Beta Function,*

$$B_x(l, m) = \int_0^x x^{l-1}(1-x)^{m-1}\,dx \quad \text{and} \quad I_x(l, m) = B_x(l, m)/B(l, m),$$

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

Solution. We have

$$I_x(l, m)\,B(l, m) = B_x(l, m) = \int_0^x x^{l-1}(1-x)^{m-1}\,dx$$

$$= \int_0^1 x^{l-1}(1-x)^{m-1}\,dx - \int_x^1 x^{l-1}(1-x)^{m-1}\,dx$$

$$= B(l, m) - \int_x^1 x^{l-1}(1-x)^{m-1}\,dx$$

In the integral, put $1 - x = y$, then

$$I_x(l, m) \cdot B(l, m) = B(l, m) - \int_{1-x}^{0} (1-y)^{l-1} y^{m-1} (-dy)$$

$$= B(l, m) - \int_0^{1-x} y^{m-1} (1-y)^{l-1} dy$$

$$= B(l, m) - B_{1-x}(m, l) = B(l, m) - I_{1-x}(m, l)\, B(m, l) \qquad \text{[From (*)]}$$

Since $B(l, m) = B(m, l)$, we get on dividing throughout by $B(l, m)$,

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

9·8. EXPONENTIAL DISTRIBUTION

Definition. *A r.v. X is said to have an exponential distribution with parameter $\theta > 0$, if its p.d.f. is given by :*

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \qquad \ldots (9\cdot26)$$

The cumulative distribution function $F(x)$ is given by

$$F(x) = \int_0^x f(u)\, du = \theta \int_0^x \exp(-\theta u)\, du \qquad \ldots (9\cdot26a)$$

$$F(x) = \begin{cases} 1 - \exp(-\theta x), \geq 0 \\ 0, \text{ otherwise} \end{cases}$$

Moment Generating Function of Exponential Distribution

$$M_X(t) = E(e^{tX}) = \theta \int_0^{\infty} e^{tx} e^{-\theta x}\, dx = \theta \int_0^{\infty} \exp\{-(\theta - t)x\}\, dx$$

$$= \frac{\theta}{(\theta - t)} = \left(1 - \frac{t}{\theta}\right)^{-1} = \sum_{r=0}^{\infty} \left(\frac{t}{\theta}\right)^r, \quad \theta > t$$

$\therefore$ $\mu_r' = E(X^r) =$ Coefficient of $\frac{t^r}{r!}$ in $M_X(t) = \frac{r!}{\theta^r}$; $r = 1, 2, \ldots$

$\therefore$ Mean $= \mu_1' = \frac{1}{\theta}$ and variance $= \mu_2 = \mu_2' - \mu_1'^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$

Hence, if $X \sim \exp(\theta)$, then Mean $= \frac{1}{\theta}$ and Variance $= \frac{1}{\theta^2}$.

Remark. Variance $= \frac{1}{\theta^2} = \frac{1}{\theta} \cdot \frac{1}{\theta} = \frac{\text{Mean}}{\theta}$

$\therefore$ Variance > Mean, if $0 < \theta < 1$

Variance = Mean, if $\theta = 1$

and Variance < Mean, if $\theta > 1$

Hence for the exponential distribution,

Variance >, =, or < Mean, for different values of the parameter.

$$= \theta \int_0^{\infty} e^{x(t-\theta)}\, dx \qquad \ldots (9\cdot26b)$$

$$= \theta \left(\frac{e^{x(t-\theta)}}{t - \theta} \right)_0^{\infty}$$

$$= \theta \left(\frac{e^{-x(\theta - t)}}{-(\theta - t)} \right)_0^{\infty}$$

$$= \theta \, \frac{e^{-\infty(\theta - t)} - \theta e^{-0}}{-(\theta - t)}$$

$$= \theta + \frac{\theta}{\theta - t}$$

$$M_X(t) = \frac{\theta}{\theta - t} = \frac{\theta}{\theta(1 - t/\theta)} = (1 - t/\theta)^{-1}$$

Theorem. *If $X_1, X_2, \ldots, X_n$ are independent r.v.'s, X_i having an exponential distribution with parameter θ_i; $i = 1, 2, \ldots, n$; then $Z = \min(X_1, X_2, \ldots, X_n)$ has exponential distribution with parameter $\sum_{i=1}^{n} \theta_i$.*

Proof. $G_Z(z) = P(Z \leq z) = 1 - P(Z > z) = 1 - P[\min(X_1, X_2, \ldots, X_n) > z]$

$$= 1 - P(X_i > z;\ i = 1, 2, \ldots, n) = 1 - \prod_{i=1}^{n} P(X_i > z)$$

($\because$ $X_1, X_2, \ldots, X_n$ are independent)

$$= 1 - \prod_{i=1}^{n} [1 - P(X_i \leq z)] = 1 - \prod_{i=1}^{n} [1 - F_{X_i}(z)]$$

(where F is the distribution function of X_i).

$$= 1 - \prod_{i=1}^{n} \left[1 - \left(1 - e^{-\theta x}\right)\right] = \begin{cases} 1 - \exp\left\{\left(-\sum_{i=1}^{n} \theta_i\right) z\right\}, & z > 0 \\ 0, \text{otherwise} \end{cases}$$

$$\therefore \quad g_Z(z) = \frac{d}{dz}(G(z)) = \begin{cases} \left(\sum_{i=1}^{n} \theta_i\right) \exp\left\{\left(-\sum_{i=1}^{n} \theta_i\right) z\right\}, & z > 0 \\ 0, \text{ otherwise} \end{cases}$$

$\Rightarrow$ $Z = \min(X_1, X_2, \ldots X_n)$ is an exponential variate with *parameter* $\sum_{i=1}^{n} \theta_i$.

Cor. If X_i; $i = 1, 2, \ldots, n$ are identically distributed, following exponential distribution with parameter θ, then $Z = \min(X_1, X_2, \ldots, X_n)$ is also exponentially distributed with parameter $n\theta$.

Example 9·41. *Show that the exponential distribution 'lacks memory', i.e., if X has an exponential distribution, then for every constant $a \geq 0$, one has*

$$P(Y \leq x \mid X \geq a) = P(X \leq x) \text{ for all } x, \text{ where } Y = X - a.$$

Solution. The *p.d.f.* of the exponential distribution with parameter θ is :

$$f(x) = \theta \exp(-\theta x);\ \theta > 0,\ 0 < x < \infty$$

We have

$$P(Y \leq x \cap X \geq a) = P(X - a \leq x \cap X \geq a) \qquad (\because Y = X - a)$$

$$= P(X \leq a + x \cap X \geq a) = P(a \leq X \leq a + x)$$

$$= \theta \int_a^{a+x} e^{-\theta x}\, dx = e^{-a\theta}(1 - e^{-\theta x})$$

and $\quad P(X \geq a) = \theta \int_a^{\infty} e^{-\theta x}\, dx = e^{-a\theta}$

$$P(Y \leq x \mid X \geq a) = \frac{P(Y \leq x \cap X \geq a)}{P(X \geq a)} = 1 - e^{-\theta x} \qquad \ldots (*)$$

Also $P(X \le x) = \theta \int_0^x e^{-\theta x}\, dx = 1 - e^{-\theta x}$ ... (**)

From (*) and (**), we get $P(Y \le x \mid X \ge a) = P(X \le x)$

i.e., *exponential distribution lacks memory.*

Example 9-42. *X and Y are independent with a common p.d.f. (exponential) :*

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

Find a *p.d.f.* for $X - Y$

Solution. Since X and Y are independent and identically distributed (*i.i.d.*), their joint *p.d.f.* is given by :

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y) = \begin{cases} e^{-(x+y)}; x > 0, y > 0 \\ 0, \text{otherwise} \end{cases}$$

Let $\begin{cases} u = x - y \\ v = y \end{cases} \Rightarrow \begin{cases} x = u + v \\ y = v \end{cases}$ ... (1)

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

Thus the joint *p.d.f.* of U and V becomes :

$$g(u, v) = e^{-(u+2v)};\ v > 0, -\infty < u < \infty$$

$(1) \Rightarrow \quad u = x - v \Rightarrow v = x - u$ Thus, $\begin{matrix} v > -u, \text{ if } -\infty < u < 0 \\ v > 0, \text{ if } u > 0 \end{matrix}$

For $-\infty < u < 0$,

$$g(u) = \int_{-u}^{\infty} g(u, v)\, dv = \int_{-u}^{\infty} e^{-(u+2v)}\, dv = e^{-u} \left| \frac{e^{-2v}}{-2} \right|_{-u}^{\infty} = \frac{1}{2} e^{u}$$

and for $u > 0$,

$$g(u) = \int_{0}^{\infty} g(u, v)\, dv = e^{-u} \left| \frac{e^{-2v}}{-2} \right|_{0}^{\infty} = \frac{1}{2} e^{-u}$$

Hence the *p.d.f.* of $U = X - Y$ is given by : $g(u) = \begin{cases} \frac{1}{2} e^{u}, & -\infty < u < 0 \\ \frac{1}{2} e^{-u}, & u > 0 \end{cases}$

These results can be combined to give :

$$g(u) = \frac{1}{2} e^{-|u|}, \quad -\infty < u < \infty$$

which is the *p.d.f.* of standard Laplace distribution (*c.f.* § 9·9).

Aliter.

$$M_X(t) = \int_0^{\infty} e^{tx} f(x)\, dx = \int_0^{\infty} e^{-(1-t)x}\, dx = \left| \frac{e^{-(1-t)x}}{-(1-t)} \right|_0^{\infty} = \frac{1}{1-t},\ t < 1$$

$\therefore$ The characteristic function of X is :

$$\varphi_X(t) = \frac{1}{1 - it} = \varphi_Y(t), \quad \text{(Since X and Y are identically distributed.)}$$

Hence $\varphi_{X-Y}(t) = \varphi_{X+(-Y)}(t) = \varphi_X(t) \cdot \varphi_{-Y}(t)$ (∵ X, Y are independent.)

$$= \varphi_X(t) \cdot \varphi_Y(-t) = \frac{1}{(1 - it)(1 + it)} = \frac{1}{1 + t^2},$$

which is the characteristic function of the Standard Laplace distribution, (*c.f.* § 9·9·1).

$$g(u) = \frac{1}{2} e^{-|u|}, \quad -\infty < u < \infty \qquad \ldots (*)$$

Hence by the uniqueness theorem of characteristic functions, $U = X - Y$ has Standard Laplace distribution with the *p.d.f.* given in (*).

9·9. STANDARD LAPLACE (DOUBLE EXPONENTIAL) DISTRIBUTION

Definition. *A continuous r.v. X is said to have standard Laplace (double exponential) distribution if its p.d.f. is given by :*

$$f(x) = \frac{1}{2} \exp.(-|x|);\ -\infty < x < \infty. \qquad \ldots (9\cdot27)$$

9·9·1. Characteristic Function of Standard Laplace Distribution.

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx = \frac{1}{2} \int_{-\infty}^{\infty} e^{itx} e^{-|x|}\, dx$$

$$= \frac{1}{2} \left[\int_{-\infty}^{\infty} \cos tx \cdot e^{-|x|}\, dx + i \int_{-\infty}^{\infty} \sin tx\, e^{-|x|}\, dx \right]$$

$$= \frac{1}{2} \cdot 2 \int_0^{\infty} \cos tx \cdot e^{-|x|}\, dx,$$

since the integrands in the first and second integrals are even and odd function of x respectively.

$$\therefore \quad \varphi_X(t) = \int_0^{\infty} e^{-x} \cos tx\, dx = 1 - t^2 \int_0^{\infty} e^{-x} \cos tx\, dx \quad \text{(On integration by parts}$$

$$= 1 - t^2\, \varphi_X(t) \qquad \ldots (9\cdot27$$

$$\Rightarrow \quad \varphi_X(t) = \frac{1}{1 + t^2}$$

$$\therefore \quad \Phi_X(t) = 1 - t^2 + t^4 - t^6 + \ldots\ldots$$

$$= 1 + 2 \cdot \frac{(it)^2}{2!} + 4! \frac{(it)^4}{4!} - \ldots$$

Hence $\kappa_1 = \kappa_3 = 0$; $\kappa_2 = 2$; $\kappa_4 = 4! = 24$

$\therefore$ Mean $= \kappa_1 = 0$; Variance $= \mu_2 = \kappa_2 = 2$

$\mu_3 = \kappa_3 = 0$; $\mu_4 = \kappa_4 + 3\kappa_2^2 = 24 + 12 = 36$

$\Rightarrow \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$; $\beta = \frac{\mu_4}{\mu_2^2} = 9$

Remark. Men deviation about mean for standard Laplace distribution is 1. (Try it).

The marginal *p.d.f.* of U is :

$$g_U(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-(1+u^2)v^2/2\right] |v|\, dv$$

$$= \frac{1}{2\pi} \cdot 2\int_0^{\infty} \exp\left[-\frac{1}{2}(1+u^2)v^2\right] |v|\, dv$$

[$\because$ Integrand is an even function of v]

$$= \frac{1}{\pi}\int_0^{\infty} \exp\left[-\frac{1}{2}(1+u^2)v^2\right] v\, dv \qquad [\because \text{For } v \geq 0,\ |v| = v]$$

$$= \frac{1}{\pi}\int_0^{\infty} e^{-t} \cdot \frac{dt}{(1+u^2)} \qquad \left[\frac{1}{2}(1+u^2)v^2 = t\right]$$

$$= \frac{1}{\pi(1+u^2)} \cdot \left|-e^{-t}\right|_0^{\infty} = \frac{1}{\pi(1+u^2)}, -\infty < u < \infty,$$

which is the *p.d.f.* of standard Cauchy distribution.

Thus the ratio of two independent standard normal variates is a standard Cauchy variate.

Example 45. Let X and Y be i.i.d. standard Cauchy variates. Prove that the p.d.f. of XY is :

$$\frac{2}{\pi^2}\left\{\frac{\log|x|}{x^2-1}\right\}.$$

Solution. Since X and Y are independent standard Cauchy variates, their joint *p.d.f.* is given by

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

Let $u = xy$ and $v = y$. Then Jacobian of transformation is given by :

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{v} & -\frac{u}{v^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{v} \qquad \left(\because y = v, x = \frac{u}{v}\right)$$

Thus the joint p.d.f. of U and V is given by :

$$g(u, v) = \frac{1}{\pi^2} \cdot \frac{1}{\left(1+\frac{u^2}{v^2}\right)(1+v^2)} \cdot \frac{1}{|v|} = \frac{1}{\pi^2} \cdot \frac{|v|}{(u^2+v^2)(1+v^2)}, -\infty < (u, v) < \infty$$

Integrating *w.r.* to v over the range $-\infty$ to ∞, marginal *p.d.f.* of U is given by :

$$g_1(u) = \int_{-\infty}^{\infty} g(u, v)\, dv = \frac{1}{\pi^2}\int_{-\infty}^{\infty} \frac{|v|}{(u^2+v^2)(1+v^2)}\, dv = \frac{2}{\pi^2}\int_0^{\infty} \frac{|v|\, dv}{(u^2+v^2)(1+v^2)}$$

(Since the integrand is an even function of v.)

$$\therefore \quad g_1(u) = \frac{2}{\pi^2}\int_0^{\infty} \frac{v}{(u^2+v^2)(1+v^2)}\, dv = \frac{1}{\pi^2}\int_0^{\infty} \frac{2v}{(u^2-1)}\left(\frac{1}{1+v^2} - \frac{1}{u^2+v^2}\right) dv$$

$$= \frac{1}{\pi^2(u^2-1)}\left|\log(1+v^2) - \log(u^2+v^2)\right|_0^{\infty} = \frac{1}{\pi^2(u^2-1)}\left|\log\left(\frac{1+v^2}{u^2+v^2}\right)\right|_0^{\infty}$$

$$= \frac{1}{\pi^2(u^2-1)}\left\{\left[\log\left(\frac{\frac{1}{v^2}+1}{\frac{u^2}{v^2}+1}\right)\right]_{v=\infty} - \log\left(\frac{1}{u^2}\right)\right\}$$

$$= \frac{1}{\pi^2(u^2-1)}\left[\log 1 + 2\log|u|\right] = \frac{2\log|u|}{\pi^2(u^2-1)}, -\infty < u < \infty$$

9.13. CENTRAL LIMIT THEOREM (C.L.T.)

The central limit theorem in the mathematical theory of probability may be expressed as follows :

"If X_i, $(i = 1, 2, \ldots, n)$ be independent random variables such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, then under certain very general conditions, the random variable $S_n = X_1 + X_2 + \ldots + X_n$, is asymptotically normal with mean μ and standard deviation σ where

$$\mu = \sum_{i=1}^{n} \mu_i \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2 \qquad \ldots(9\cdot32)$$

This theorem was first stated by Laplace in 1812 and a regorous proof under fairly general conditions was given by Liapounoff in 1901. Below we shall consider some particular cases of this general central limit theorem.

9.13.1. De-Moivre's Laplace Theorem (1733). A particular case of central limit theorem is De-Moivre's theorem which states as follows :

"If

$$X_i = \begin{cases} 1, \text{ with probability } p \\ 0, \text{ with probability } q \end{cases}$$

then the distribution of the random variable $S_n = X_1 + X_2 + \ldots + X_n$, where X_i's are independent, is asymptotically normal as $n \to \infty$."

Proof. M.G.F. of X_i is given by : $M_{X_i}(t) = E(e^{tX}) = e^{t\cdot 1}p + e^{t\cdot 0}q = (q + pe^t)$ ... (*)

M.G.F. of the sum $S_n = X_1 + X_2 + \ldots + X_n$ is given by

$$M_{S_n}(t) = M_{X_i + X_2 + \ldots + X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \ldots M_{X_n}(t) = [M_{X_i}(t)]^n$$

(since X_i's are independently and identically distributed.)

$$= (q + pe^t)^n, \qquad \text{[From (*)]} \qquad \ldots (**)$$

which is the M.G.F. of a binomial variate with parameters n and p.

Hence by uniqueness theorem of *m.g.f.*'s, $S_n \sim B(n, p)$.

$\therefore$ $E(S_n) = np = \mu$ (say), and $V(S_n) = npq = \sigma^2$, (say).

Let $$Z = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - \mu}{\sigma}$$

$$M_Z(t) = e^{-\mu t/\sigma} M_{S_n}(t/\sigma) = e^{-np t/\sqrt{npq}}\left[q + pe^{t/\sqrt{npq}}\right]^n \qquad \text{[From (**)]}$$

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

where $O(n^{-\frac{3}{2}})$ represents terms involving $n^{3/2}$ and higher powers of n in the denominator.

Proceeding to the limits as $n \to \infty$, we get

$$\lim_{n\to\infty} M_Z(t) = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + O(n^{-3/2})\right]^n = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n}\right]^n = e^{t^2/2}, \quad (61)$$

which is the M.G.F. of a standard normal variate. Hence by the uniqueness theorem of M.G.F.'s, $Z = \frac{S_n - \mu}{\sigma}$ is asymptotically $N(0, 1)$.

Hence $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically $N(\mu, \sigma^2)$ as $n \to \infty$.

Remarks **1.** From this theorem it follows that standard binomial variate tends to standard normal variate as $n \to \infty$. In other words, *binomial distribution tends to normal distribution as* $n \to \infty$.

2. *Convergence in Distribution or Law.* Let $\{X_n\}$ be a sequence of r.v.'s and $\{F_n\}$ be the corresponding sequence of distribution functions. We say that X_n converges in distribution (or law) to X if there exists a r.v. X with distribution function F such that as $n \to \infty$, $F_n(x) \to F(x)$ at every point x at which F is continuous.

We write $X_n \xrightarrow{L} X$ or $X_n \xrightarrow{d} X$.

3. It may be remarked that convergence in probability discussed in Chapter 7 implies convergence in distribution (or law), *i.e.*, $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X$. ... (*)

The converse is not true, *i.e.*, $X_n \xrightarrow{P} X$, in general, does not imply $X_n \xrightarrow{P} X$.

However, we have the following result. Let k be a constant. Then $X_n \xrightarrow{L} k \Rightarrow X_n \xrightarrow{P} k$ ... (**)

Combining (*) and (**), we get the following result.

Let k be a constant. Then $X_n \xrightarrow{L} k \Leftrightarrow X_n \xrightarrow{P} k$ ... (***)

9·13·2. Lindeberg-Levy Theorem. The following case of central limit theorem for independent and identically distributed (*i.i.d.*) variables, was first proved by Lindeberg and Levy.

> *"If $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables with*
>
> $$\left.\begin{aligned} E(X_i) &= \mu_1 \\ V(X_i) &= \sigma_1^2 \end{aligned}\right\}, \; i = 1, 2, \ldots, n$$
>
> *then the sum $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically normal with mean $\mu = n\mu_1$ and variance $\sigma^2 = n\sigma_1^2$."*

Here we make the following assumptions :

(*i*) The variables are independent and identically distributed.

(*ii*) $E(X_i^2)$ exists for all $i = 1, 2, 3, \ldots$

Proof. Let $M_1(t)$ denote the m.g.f. of each of the deviation $(X_i - \mu_1)$ and $M(t)$ denote the m.g.f. of the standard variate $Z = (S_n - \mu)/\sigma$.

Since μ_1' and μ_2', (about origin) of the deviation $(X_i - \mu_1)$ are given by :

$$\mu_1' = E(X_i - \mu_1) = 0, \qquad \mu_2' = E(X_i - \mu_1)^2 = \sigma_1^2$$

We have

$$M_1(t) = \left(1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \ldots\right) = \left\{1 + \frac{t^2}{2!}\sigma_1^2 + O(t^3)\right\}, \quad \ldots (*)$$

where $O(t^3)$ contains terms with t^3 and higher powers of t.

We have $Z = \frac{S_n - \mu}{\sigma} = \frac{(X_1 + X_2 + \ldots + X_n) - n\mu_1}{\sigma} = \sum_{i=1}^{n}\left(\frac{X_i - \mu_1}{\sigma}\right)$ (62)

and since X_i's are independent, we get

$$M_Z(t) = M \sum_{i=1}^{n} (X_i - \mu_1)/\sigma (t) = M \sum_{i=1}^{n} (X_i - \mu_1)(t/\sigma)$$

$$= \prod_{i=1}^{n} [M_{(X_i - \mu_1)}(t/\sigma)] = [M_1(t/\sigma)]^n = [M_1(t/\sqrt{n}\,\sigma_1)]^n$$

$$= \left[1 + \frac{t^2}{2n} + O(n^{-3/2})\right]^n \qquad \text{[From (*)]}$$

For every fixed 't', the terms $O(n^{-3/2}) \to 0$ as $n \to \infty$. Therefore, as $n \to \infty$, we get

$$\lim_{n\to\infty} M_Z(t) = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + O(n^{-3/2})\right]^n = \exp(t^2/2),$$

which is the M.G.F. of standard normal variate. Hence by uniqueness theorem of M.G.F.'s $Z = (S_n - \mu)/\sigma$ is asymptotically $N(0, 1)$, or $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically $N(\mu, \sigma^2)$, where $\mu = n\mu_1$ and $\sigma^2 = n\sigma_1^2$.

Note. C.L.T. can be stated in another from as follows :

(*i*) If $X_1, X_2, \ldots, X_n$ are *i.i.d.* with mean μ_1 and variance σ_1^2 (finite) and $S_n = X_1 + X_2 + \ldots + X_n$, then for $-\infty < a < b < \infty$;

$$\lim_{n\to\infty} P\left[a \le \frac{S_n - n\mu_1}{\sigma_1\sqrt{n}} \le b\right] = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \quad \ldots (9\cdot33)$$

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

or (*ii*) $$\lim_{n\to\infty} P\left[a \le \frac{S_n - E(S_n)}{\sqrt{\text{Var}\,(S_n)}} \le b\right] = \Phi(b) - \Phi(a) \quad \ldots (9\cdot33a)$$

or, still another form :

(*iii*) $$\lim_{n\to\infty} P\left[a \le \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\text{Var}\,(\bar{X}_n)}} \le b\right] = \Phi(b) - \Phi(a)$$

i.e., $$\lim_{n\to\infty} P\left[a \le \frac{\bar{X}_n - \mu_1}{\sigma_1/\sqrt{n}} \le b\right] = \Phi(b) - \Phi(a) \quad \ldots (9\cdot33b)$$

Remarks **1.** We wrote the C.L.T. using non-strict inequalities $P[a \le (\cdot) \le b]$

It makes no difference whether one or both are changed to a strict inequality. The reason is that the limit distribution function (*d.f.*) $\Phi(\cdot)$ is a continuous *d.f.*

2. In the binomial case, C.L.T. gives good approximation if p is nearly 1/2. For p near about 0 or 1, C.L.T. approximation still holds but in that case n has to be sufficiently greater than in the case $p = 1/2$ approximately.

9·13·3. Applications of Central Limit Theorem. (*a*) If $X_1, X_2, \ldots$ are i.i.d. B (r,

and $S_n = X_1 + X_2 + \ldots + X_n$, then $E(S_n) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} (rp) = nrp$

and $V(S_n) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} (rpq) = nrpq,$ $(q = 1$

Hence (9·33a),

$$\Rightarrow \lim_{n\to\infty} P\left[a \le \frac{S_n - nrp}{\sqrt{nrp(1-p)}} \le b\right] = \Phi(b) - \Phi(a),\ 0 < p < 1$$

(b) If Y_n is binomial variate with parameters n and p, then

$$\lim_{n\to\infty} P\left[a \le \frac{Y_n - np}{\sqrt{npq}(1-p)} \le b\right] = \Phi(b) - \Phi(a),\ 0 < p < 1$$

Proof. Let $X_1, X_2, \ldots$ be *i.i.d.* Bernoulli variates, *i.e.*, $B(1, p)$, then

$$S_n = X_1 + X_2 + \ldots + X_n = B(n, p). \text{ But } Y_n = B(n, p)$$

Hence using Y_n instead of S_n in (9·33a), we get

$$\lim_{n\to\infty} P\left[a \le \frac{Y_n - E(Y_n)}{\sqrt{Var\ Y_n}} \le b\right] = \Phi(b) - \Phi(a)$$

i.e.,

$$\lim_{n\to\infty} P\left[a \le \frac{Y_n - np}{\sqrt{npq}} \le b\right] = \Phi(b) - \Phi(a),\ q = 1 - p$$

(c) If Y_n is distributed as $P(n)$, then $\lim_{n\to\infty} P\left[a \le \frac{Y_n - n}{\sqrt{n}} \le b\right] = \Phi(b) - \Phi(a)$

Thus, for instance $\lim_{n\to\infty} P(Y_n \le n) = \frac{1}{2}$, (*i.e.*,) $\sum_{k=0}^{n} \frac{e^{-n} n^k}{k!} = \frac{1}{2}$, as $n \to \infty$

Proof. Let $X_1, X_2, \ldots$ be *i.i.d.* $P(1)$. Then $S_n = X_1 + X_2 + \ldots + X_n \sim P(n) \Rightarrow Y_n = S_n$

$$\therefore \quad P\left[a \le \frac{Y_n - n}{\sqrt{n}} \le b\right] = P\left[a \le \frac{S_n - n}{\sqrt{n}} \le b\right] \to \Phi(b) - \Phi(a) \text{ as } n \to \infty$$

In particular, let us take $a = -\infty$ and $b = 0$, then

$$P\left(a \le \frac{Y_n - n}{\sqrt{n}} \le b\right) = P\left(\frac{Y_n - n}{\sqrt{n}} \le 0\right) = P(Y_n \le n) \quad \ldots (*)$$

Also $\Phi(b) - \Phi(a) = \Phi(0) - \Phi(-\infty) = \frac{1}{2}$ $\quad \ldots (**)$

From (*) and (**), we get : $P(Y_n \le n) \to \frac{1}{2}$ as $n \to \infty$.

Remark. This result could be generalised. In fact, on taking $a = -\infty$ and $b = 0$ in (9·33a), we get

$$P\left[\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \le 0\right] \to \frac{1}{2} \quad \Rightarrow \quad P[S_n \le E(S_n)] \to \frac{1}{2}, \text{ as } n \to \infty$$

9·13·4. Liapounoff's Central Limit Theorem. Below we shall give (without proof) the central limit theorem for the generalised case when the variables are not identically distributed and where, in addition to the existence of the second moment for the variables X_i, we impose some further conditions.

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that

$$\left.\begin{aligned} E(X_i) &= \mu_i \\ V(X_i) &= \sigma_i^2 \end{aligned}\right\};\ i = 1, 2, \ldots, n.$$

Let us suppose that third absolute moment, say ρ_i^3 of X_i about its mean exists

i.e., $\rho_i^3 = E(|X_i - \mu_i|^3);\ i = 1, 2, \ldots, n$ is finite.

Further let $\rho^3 = \sum_{i=1}^{n} \rho_i^3$.

If $\lim_{n\to\infty} \frac{\rho}{\sigma} = 0$, the sum $X = X_1 + X_2 + \ldots + X_n$ is asymptotically $N(\mu, \sigma^2)$,

where $\mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$

Remarks on Central Limit Theorems.

(About Liapounoff's theorem). If the variables $X_i;\ i = 1, 2, \ldots, n$ are identical, then

$$\rho^3 = \sum_{i=1}^{n} \rho_i^3 = n\rho_1^3 \text{ and } \sigma^2 = \sum_{i=1}^{n} \sigma_i^2 = n\sigma_1^2, \quad \therefore \frac{\rho}{\sigma} = \frac{n^{1/3}\rho_1}{n^{1/2}\cdot\sigma_1} = \frac{\rho_1}{\sigma_1}\cdot\frac{1}{n^{1/6}} \to 0, \text{ as } n \to \infty.$$

Thus for identical variables, the condition of Liapounoff's theorem is satisfied.

It may be pointed out here that Lindeberg - Levy theorem proved in § 9·13·2, should not be inferred as a particular case of Liapounoff's theorem, since the former does not assume the existence of the third moment.

2. Central limit theorem can be expected in the following cases :

(*i*) If a certain random variable X arises as cumulative effect of several independent causes each of which can be considered as a continuous random variable, then X obeys central limit theorem under certain regularity conditions.

(*ii*) If $\varphi(X_1, X_2, \ldots, X_n)$, is a function of X_i's having first and second continuous derivatives about the point $(\mu_1, \mu_2, \ldots, \mu_n)$, then under certain regularity conditions, $\varphi(X_1, X_2, \ldots, X_n)$ is asymptotically normal with mean $\varphi(\mu_1, \mu_2, \ldots, \mu_n)$.

(*iii*) Under certain conditions, the central limit theorem holds for variables which are not independent.

3. *Relation between CLT and WLLN :*

(*a*) Both the central limit theorem (CLT) and the weak law of large numbers (WLLN) hold for a sequence $\{X_n\}$ of *i.i.d.* random variables with finite mean μ and variance σ^2.

However, in this case, the CLT is a stronger result than the WLLN in the sense that the former provides an estimate of the $P[|(S_n - n\mu)/n| \ge \varepsilon]$, as given below :

$$P\left[\left|\frac{S_n - n\mu}{n}\right| \ge \varepsilon\right] = P\left[\left|\overline{X}_n - \mu\right| \ge \varepsilon\right] = P\left[\left|\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right| \ge \frac{\varepsilon}{\sigma/\sqrt{n}}\right]$$

$$= P[|Z| \ge \varepsilon\sqrt{n}/\sigma];\ Z \sim N(0, 1)$$

$$= 1 - P[|Z| \le \varepsilon\sqrt{n}/\sigma] = 1 - \left[\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right]$$

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

However, WLLN does not require the existence of variance (*c.f. Khinchine's Theorem*).

(*b*) For the sequence $\{X_n\}$ of independent and uniformly bounded r.v.'s, WLLN holds and CLT holds in this case provided $B_n = Var(X_1 + X_2 + \ldots + X_n) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2 \to \infty$ as $n \to \infty$.

(*c*) For the sequence $\{X_n\}$ of independent r.v.'s, CLT may hold but the WLLN may not hold.

9·13·5. Cramer's Theorem. We state below (without proof), a useful result on the convergence of sequences of r.v.'s.

Cramer's Theorem. Let $\{X_n\}$ and $\{Y_n\}$ be sequences of r.v.'s such that

$X_n \xrightarrow{L} X$ and $Y_n \xrightarrow{P} c$ (constant), then $\frac{X_n}{Y_n} \xrightarrow{L} \frac{X}{c}$ if $c \ne 0$

7·7. WEAK LAW OF LARGE NUMBERS
7·8. BOREL-CANTELLI LEMMA
7·9. PROBABILITY GENERATING FUNCTION
7·9·1. Convolutions.

CHAPTER CONCEPTS QUIZ / DISCUSSION & REVIEW QUESTIONS / ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

65

7·1. MOMENT GENERATING FUNCTION

The moment generating function (*m.g.f.*) of a random variable X (about origin) having the probability function $f(x)$ is given by :

$$M_X(t) = E(e^{tX}) = \begin{cases} \int e^{tx} f(x)\,dx, & \text{(for continuous probability distribution)} \\ \sum_x e^{tx} f(x), & \text{(for discrete probability distribution)} \end{cases} \quad \ldots (7\cdot1)$$

the integration or summation being extended to the entire range of x, t being the real parameter and it is being assumed that the right-hand side of (7·1) is absolutely convergent for some positive number h such that $-h < t < h$. Thus

$$M_X(t) = E(e^{tX}) = E\left(1 + tX + \frac{t^2X^2}{2!} + \ldots + \frac{t^rX^r}{r!} + \ldots\right)$$

$$= 1 + t\,E(X) + \frac{t^2}{2!}E(X^2) + \ldots + \frac{t^r}{r!}E(X^r) + \ldots$$

$$= 1 + t\,\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots \quad \ldots (7\cdot2)$$

$$= \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r' \quad \ldots (7\cdot2a)$$

$$\mu_r' = E(X^r) = \begin{cases} \int x^r f(x)\,dx, & \text{for continuous distribution} \\ \sum_x x^r p(x)\,dx, & \text{for discrete distribution,} \end{cases}$$

the rth moment of X about origin. Thus we see that the coefficient of $\frac{t^r}{r!}$ in $M_X(t)$ is μ_r' (about origin). Since $M_X(t)$ generates moments, it is known as *moment generating function*.

Differentiating (7·2) w.r. to. t r times and then putting $t = 0$, we get

$$\left|\frac{d^r}{dt^r}(M_X(t))\right|_{t=0} = \left|\frac{\mu_r'}{r!}\cdot r! + \mu_{r+1}'\,t + \mu_{r+2}'\cdot\frac{t^2}{2!} + \ldots\right|_{t=0}$$

$$\mu_r' = \left|\frac{d^r}{dt^r}(M_X(t))\right|_{t=0} \quad \ldots (7\cdot3)$$

In general, the moment generating function X about the point $X = a$ is defined as :

66

$$M_X(t) \text{ (about } X = a) = E[e^{t(X-a)}]$$

$$= E\left\{1 + t(X-a) + \frac{t^2}{2!}(X-a)^2 + \ldots + \frac{t^r}{r!}(X-a)^r + \ldots\right\}$$

$$= 1 + t\,\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots \quad \ldots (7\cdot4)$$

where $\mu_r' = E[(X-a)^r]$, is the rth moment about the point $X = a$.

7·1·1. Some Limitations of Moment Generating Function. Moment generating function suffers from some drawbacks which have restricted its use in Statistics. We give below some of the deficiencies of *m.g.f.'s* with illustrative examples.

1. *A random variable X may have no moments although its m.g.f. exists.* For example, let us consider a discrete *r.v.* with probability function :

$$f(x) = \begin{cases} \frac{1}{x(x+1)}; x = 1, 2, \ldots \\ 0, \text{ otherwise} \end{cases}$$

Here $E(X) = \sum_{x=1}^{\infty} x f(x) = \sum_{x=1}^{\infty} \frac{1}{(x+1)} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots = \left(\sum_{x=1}^{\infty} \frac{1}{x}\right) - 1$

Since, $\sum_{x=1}^{\infty} \frac{1}{x}$ is a divergent series, $E(X)$ does not exist and consequently no moment of X exists. However, the *m.g.f.* of X is given by :

$$M_X(t) = \sum_{x=1}^{\infty} e^{tx} f(x) = \sum_{x=1}^{\infty} \frac{e^{tx}}{x(x+1)} = \sum_{x=1}^{\infty} \frac{z^x}{x(x+1)}, \quad (z = e^t) \quad \ldots (*)$$

$$= \frac{z}{1\cdot2} + \frac{z^2}{2\cdot3} + \frac{z^3}{3\cdot4} + \frac{z^4}{4\cdot5} + \ldots = z\left(1 - \frac{1}{2}\right) + z^2\left(\frac{1}{2} - \frac{1}{3}\right) + z^3\left(\frac{1}{3} - \frac{1}{4}\right) + z^4\left(\frac{1}{4} - \frac{1}{5}\right) +$$

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

$$= -\log(1-z) - \frac{1}{z}\left(\frac{z^2}{2} + \frac{z^3}{3} + \frac{z^4}{4} + \ldots\right) = -\log(1-z) + 1 + \frac{1}{z}\log(1-z), \; |z|$$

$$= 1 + \left(\frac{1}{z} - 1\right)\log(1-z), \; |z| < 1$$

$$= 1 + (e^{-t} - 1)\log(1 - e^t), \; t < 0 \quad \text{[From (*)]} \qquad [\because |z| < 1 \Rightarrow |e^t| < 1 \Rightarrow t <$$

Also $M_X(t) = 1$ for $t = 0$.

Hence $M_X(t)$ exists for $t \leq 0$.

2. *A random variable X can have m.g.f. and some (or all) moments, yet the m.g.f. do generate the moments. For example, consider a discrete r.v. with probability function :*

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

$$E(X^r) = \sum_{x=0}^{\infty} (2^x)^r\, P(X = 2^x) = e^{-1} \sum_{x=0}^{\infty} \frac{(2^r)^x}{x!} = e^{-1} \cdot \exp(2^r) = \exp(2^r - 1)$$

Hence all the moments of X exist. The m.g.f. of X, if it exists, is given by :

$$M_X(t) = \sum_{x=0}^{\infty} \exp(t \cdot 2^x)\left(\frac{e^{-1}}{x!}\right) = e^{-1} \sum_{x=0}^{\infty} \exp(t \cdot 2^x)\frac{1}{x!}$$

By D' Alembert's ratio test, the series on the R.H.S. converges for $t \leq 0$ and diverges for $t > 0$. Hence $M_X(t)$ cannot be differentiated at $t = 0$ and has no Maclaurin's expansion and consequently it does not generate moments.

3. *A r.v. X can have all or some moments, but m.g.f. does not exist except perhaps at one point.*

For example, let X be a r.v. with probability function :

$$P(X = \pm 2^x) = \begin{cases} \frac{e^{-1}}{2x!}; x = 0, 1, 2, \ldots \\ 0, \text{ otherwise} \end{cases}$$

Since the distribution is symmetric about the line $X = 0$, all moments of odd order about origin vanish, *i.e.*, $E(X^{2r+1}) = 0 \Rightarrow \mu_{2r+1} = 0$.

$$E(X^{2r}) = \sum_{x=0}^{\infty} (\pm 2^x)^{2r}\left(\frac{1}{2ex!}\right) = \frac{1}{e}\sum_{x=0}^{\infty}\frac{(2^{2r})^x}{x!} = \frac{1}{e}\exp(2^{2r}) = \exp(2^{2r}-1)$$

Thus all the moments of X exist. The m.g.f. of X, if it exists, is given by

$$M_X(t) = \sum_{x=0}^{\infty}\left\{(e^{t\cdot 2^x} + e^{-t\cdot 2^x})\frac{1}{2ex!}\right\} = e^{-1}\sum_{x=0}^{\infty}\left\{\frac{\cosh(t2^x)}{x!}\right\},$$

which converges only for $t = 0$.

As an illustration of a continuous probability distribution, consider Pareto distribution with p.d.f. : $p(x) = \frac{\theta a^\theta}{x^{\theta+1}}; x \geq a, \theta > 1$.

$$E(X^r) = \theta a^\theta \int_a^\infty x^{r-\theta-1}\,dx = \theta a^\theta \left|\frac{x^{r-\theta}}{r-\theta}\right|_a^\infty, \text{ which is finite iff } r - \theta < 0 \Rightarrow \theta > r$$

and then $$E(X^r) = \theta a^\theta\left[0 - \frac{a^{r-\theta}}{r-\theta}\right] = \frac{\theta \cdot a^r}{\theta - r};\ \theta > r$$

However, the *m.g.f.* is given by :

$M_X(t) = \theta a^\theta \int_a^\infty \frac{e^{tx}}{x^{\theta+1}}\,dx$, which does not exist, since e^{tx} dominates $x^{\theta+1}$ and $(e^{tx}/x^{\theta+1}) \to \infty$ as $x \to \infty$ and hence the integral is not convergent.

For more illustrations see Student's *t*-distribution and Snedecor's *F*-distribution, which m.g.f.'s do not exist, though moments of all orders exist. [*c.f.* Chapter 15].

Remark. The reason that *m.g.f.* is a poor tool in comparison with characteristic function (§ 7·3) is that the domain of the dummy parameter '*t*' of the *m.g.f.* depends on the distribution of the r.v. under consideration, while characteristic function exists for all real t, $(-\infty < t < \infty)$. If *m.g.f.* is valid for t lying in an interval containing zero, then *m.g.f.* can be extended with perhaps some additional restrictions.

7·1·2. Properties of Moment Generating Function

Property 1. *$M_{cX}(t) = M_X(ct)$, c being a constant.* ... (7·5)

Proof. By *def.*, L.H.S. $= M_{cX}(t) = E(e^{tcX})$, R.H.S. $= M_X(ct) = E(e^{ctX})$ = L.H.S.

Property 2. *The moment generating function of the sum of a number of independent variables is equal to the product of their respective moment generating functions.*

Symbolically, if $X_1, X_2, \ldots, X_n$ are independent random variables, then the moment generating function of their sum $X_1 + X_2 + \ldots + X_n$ is given by :

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t)\,M_{X_2}(t)\ldots M_{X_n}(t)$$

Proof. By def.,

$$M_{X_1+X_2+\ldots+X_n}(t) = E[e^{t(X_1+X_2+\ldots+X_n)}] = E[e^{tX_1}\cdot e^{tX_2}\ldots e^{tX_n}]$$
$$= E(e^{tX_1})\cdot E(e^{tX_2})\ldots E(e^{tX_n})$$
$$(\because X_1, X_2, \ldots, X_n \text{ are indepen}$$
$$= M_{X_1}(t)\,M_{X_2}(t)\ldots M_{X_n}(t).$$

Property 3. *Effect of change of origin and scale on M.G.F.* Let us transform X to new variable U by changing both the origin and scale in X as follows :

$U = \frac{X-a}{h}$, where a and h are constants.

M.G.F. of U (about origin) is given by :

$$M_U(t) = E(e^{tU}) = E[\exp\{t(x-a)/h\}] = E[e^{tX/h}\,e^{-at/h}] = e^{-at/h}E(e^{tX/h})$$
$$= e^{-at/h}E(e^{Xt/h}) = e^{-at/h}M_X(t/h),$$

where $M_X(t)$ is the *m.g.f.* of X about origin.

In particular, if we take $a = E(X) = \mu$, (say), and $h = \sigma_X = \sigma$, (say), then

$$U = \frac{X - E(X)}{\sigma_X} = \frac{X-\mu}{\sigma} = Z, \text{ (say)},$$

is known as a *standard variate*. Thus the *m.g.f.* of a standard variate Z is given by

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

Remark $E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}[E(X)-\mu] = \frac{1}{\sigma}(\mu-\mu) = 0$

and $V(Z) = V\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(X-\mu) = \frac{1}{\sigma^2}V(X) = \frac{1}{\sigma^2}\cdot\sigma^2 = 1.$

$\therefore E(Z) = 0$ and $V(Z) = 1$, *i.e.*, mean and variance of a standard variate are 0 and 1 respectively.

7·1·3. Uniqueness Theorem of Moment Generating Function. *The moment generating function of a distribution, if it exists, uniquelly determines the distribution* i.e., implies that corresponding to a given probability distribution, there is only one m.g.f. (provided it exists) and corresponding to a given *m.g.f.*, there is only one probability distribution. Hence $M_X(t) = M_Y(t) \Rightarrow X$ and Y are identically distributed. [For further discussion, see Uniqueness Theorem of Characteristic Functions.]

7·2. CUMULANTS

Cumulants generating function $K(t)$ is defined as : $K_X(t) = \log_e M_X(t)$, provided the right-hand side can be expanded as a convergent series in powers of t. Thus

$$K_X(t) = \kappa_1 t + \kappa_2\frac{t^2}{2!} + \ldots + \kappa_r\frac{t^r}{r!} + \ldots = \log M_X(t)$$
$$= \log\left(1 + \mu_1' t + \mu_2'\frac{t^2}{2!} + \mu_3'\frac{t^3}{3!} + \ldots + \mu_r'\frac{t^r}{r!} + \ldots\right)$$

where κ_r = coefficient of $\frac{t^r}{r!}$ in $K_X(t)$ is called the rth cumulant. Hence

(69)

$$\kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \kappa_4 \frac{t^4}{4!} + \ldots$$

$$= \left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \ldots\right) - \frac{1}{2}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \ldots\right)^2$$

$$+ \frac{1}{3}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \ldots\right)^3 - \frac{1}{4}\left(\mu_1' t + \mu_2' \frac{t^2}{2!} + \ldots\right)^4 + \ldots$$

Comparing the coefficients of like powers of 't' on both sides, we get the relationship between the moments and cumulants. Hence, we have

$$\kappa_1 = \mu_1' = \text{Mean}, \quad \frac{\kappa_2}{2!} = \frac{\mu_2'}{2!} - \frac{\mu_1'^2}{2!} \Rightarrow \kappa_2 = \mu_2' - \mu_1'^2 = \mu_2$$

$$\frac{\kappa_3}{3!} = \frac{\mu_3'}{3!} - \frac{1}{2}\cdot\frac{2\mu_1'\mu_2'}{2!} + \frac{\mu_1'^3}{3} \Rightarrow \kappa_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \mu_3$$

Also $$\frac{\kappa_4}{4!} = \frac{\mu_4'}{4} - \frac{1}{2}\left(\frac{\mu_2'^2}{4} + \frac{2\mu_1'\mu_3'}{3!}\right) + \frac{1}{3}\cdot\frac{3\mu_1'^2\mu_2'}{2} - \frac{\mu_1'^4}{4}$$

$$\Rightarrow \quad \kappa_4 = \mu_4' - 3\mu_2'^2 - 4\mu_1'\mu_3' + 12\mu_1'^2\mu_2' - 6\mu_1'^4$$

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

$$= \mu_4 - 3(\mu_2' - \mu_1'^2)^2 = \mu_4 - 3\mu_2^2 = \mu_4 - 3\kappa_2^2 \qquad (\because \mu_2 = \kappa_2)$$

$$\therefore \quad \mu_4 = \kappa_4 + 3\kappa_2^2$$

Hence we have obtained :

$$\left.\begin{aligned} \text{Mean} &= \kappa_1 \\ \mu_2 &= \kappa_2 = \text{Variance} \\ \mu_3 &= \kappa_3 \\ \mu_4 &= \kappa_4 + 3\kappa_2^2 \end{aligned}\right\} \quad \ldots (7\cdot9b)$$

Remarks 1. These results are of fundamental importance and should be committed to memory.

2. If we differentiate both sides of (7·9a) w.r. to t 'r' times and then put $t = 0$, we get :

$$\kappa_r = \left|\frac{d^r}{dt^r} K_X(t)\right|_{t=0} \quad \ldots (7\cdot9c)$$

1. PROPERTIES OF CUMULANTS

Property 1. ***Additive Property of Cumulants.*** *The rth cumulant of the sum of the independent random variables is equal to the sum of the rth cumulants of the individual variables. Symbolically,*

$$\kappa_r(X_1 + X_2 + \ldots + X_n) = \kappa_r(X_1) + \kappa_r(X_2) + \ldots + \kappa_r(X_n), \quad \ldots (7\cdot10)$$

where X_i ; $i = 1, 2, \ldots, n$ are independent random variables.

Proof. Since X_i's are independent, $M_{X_1 + X_2 + \ldots + X_n}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_n}(t)$

Taking logarithm of each side, $K_{X_1 + X_2 + \ldots + X_n}(t) = K_{X_1}(t) + K_{X_2}(t) + \ldots + K_{X_n}(t)$

Differentiating both sides w.r.to. t 'r' times and putting $t = 0$, we get

$$\left.\frac{d^r}{dt^r} K_{X_1 + X_2 + \ldots + X_n}(t)\right|_{t=0} = \left[\frac{d^r}{dt^r} K_{X_1}(t)\right]_{t=0} + \left[\frac{d^r}{dt^r} K_{X_2}(t)\right]_{t=0} + \ldots + \left[\frac{d^r}{dt^r} K_{X_n}(t)\right]_{t=0}$$

$$\kappa_r(X_1 + X_2 + \ldots + X_n) = \kappa_r(X_1) + \kappa_r(X_2) + \ldots + \kappa_r(X_n).$$

(70)

Property 2. ***Effect of Change of Origin and Scale on Cumulants.*** If we take $U = \frac{X - a}{h}$, then $M_U(t) = \exp(-at/h)\, M_X(t/h)$

$$\therefore \quad K_U(t) = \log M_U(t) = -\frac{at}{h} + K_X(t/h)$$

$$\kappa_1' t + \kappa_2' \frac{t^2}{2!} + \ldots + \kappa_r' \frac{t^r}{r!} + \ldots = -\frac{at}{h} + \kappa_1(t/h) + \kappa_2 \frac{(t/h)^2}{2!} + \ldots + \kappa_r \frac{(t/h)^r}{r!} + \ldots$$

where κ_r' and κ_r are the rth cumulants of U and X respectively.

Comparing coefficients, we get $\kappa_1' = \frac{\kappa_1 - a}{h}$ and $\kappa_r' = \frac{\kappa_r}{h^r}$; $r = 2, 3, \ldots$ (7·10a)

Thus we see that except the first cumulant, all cumulants are independent of change of origin. But the cumulants are not invariant of change of scale as the rth cumulant of U is $(1/h^r)$ times the rth cumulant of the distribution of X.

Example 7·1. *Let the r.v. X assume the value 'r' with the probability law :*

$P(X = r) = q^{r-1} p$; $r = 1, 2, 3, \ldots$ *Find the m.g.f. of X and hence its mean and variance.*

Solution. $$M_X(t) = E(e^{tX}) = \sum_{r=1}^{\infty} e^{tr} q^{r-1} p = \frac{p}{q} \sum_{r=1}^{\infty} (qe^t)^r$$

$$= \frac{p}{q} qe^t \sum_{r=1}^{\infty} (qe^t)^{r-1} = pe^t \left[1 + qe^t + (qe^t)^2 + \ldots\right] = \left(\frac{pe^t}{1 - qe^t}\right)$$

If dash (') denotes the differentiation w.r. to t, then we have

$$M_X'(t) = \frac{pe^t}{(1 - qe^t)^2}, \qquad M_X''(t) = pe^t \frac{(1 + qe^t)}{(1 - qe^t)^3}$$

$$\therefore \quad \mu_1' \text{ (about origin)} = M_X'(0) = \frac{p}{(1-q)^2} = \frac{1}{p}$$

$$\mu_2' \text{ (about origin)} = M_X''(0) = \frac{p(1+q)}{(1-q)^3} = \frac{1+q}{p^2}.$$

Hence, mean $= \mu_1'$ (about origin) $= \frac{1}{p}$ and variance $= \mu_2 = \mu_2' - \mu_1'^2 = \frac{1+q}{p^2} -$

Example 7·2. *The probability density function of the random variable X fo probability law :*

$$p(x) = \frac{1}{2\theta} \exp\left(-\frac{|x - \theta|}{\theta}\right), \; -\infty < x < \infty$$

Find M.G.F. of X. Hence or otherwise find E (X) and V (X).

Solution. The moment generating function of X is :

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{2\theta} \exp\left(-\frac{|x - \theta|}{\theta}\right) e^{tx}\, dx$$

$$= \int_{-\infty}^{\theta} \frac{1}{2\theta} \exp\left(-\frac{\theta - x}{\theta}\right) e^{tx}\, dx + \int_{\theta}^{\infty} \frac{1}{2\theta} \exp\left(-\frac{x - \theta}{\theta}\right)$$

[For $x \in (-\infty, \theta)$, $x - \theta < 0 \Rightarrow \theta - x > 0$

$\therefore \quad |x - \theta| = \theta - x \; \forall \; x \in (-\infty, \theta)$. Similarly, $|x - \theta| = x - \theta \; \forall \; x \in (\theta,$

$$\therefore \quad M_X(t) = \frac{e^{-1}}{2\theta} \int_{-\infty}^{\theta} \exp\left\{x\left(t + \frac{1}{\theta}\right)\right\} dx + \frac{e}{2\theta} \int_{\theta}^{\infty} \exp\left\{-x\left(\frac{1}{\theta} - t\right)\right.$$

$$= \frac{e^{-1}}{2\theta}\,\frac{1}{\left(t+\frac{1}{\theta}\right)}\cdot\exp\left\{\theta\left(t+\frac{1}{\theta}\right)\right\} + \frac{e}{2\theta}\cdot\frac{1}{\left(\frac{1}{\theta}-t\right)}\cdot\exp\left\{-\theta\left(\frac{1}{\theta}-t\right)\right\}$$

$$= \frac{e^{\theta t}}{2(\theta t+1)} + \frac{e^{\theta t}}{2(1-\theta t)} = \frac{e^{\theta t}}{1-\theta^2 t^2} = e^{\theta t}\,(1-\theta^2 t^2)^{-1}$$

$$= \left(1+\theta t+\frac{\theta^2 t^2}{2!}+\ldots\right)(1+\theta^2 t^2+\theta^4 t^4+\ldots) = 1+\theta t+\frac{3\theta^2 t^2}{2!}+\ldots$$

$\therefore$ $E(X) = \mu_1' =$ Coeff. of t in $M_X(t) = \theta$, $\quad \mu_2' =$ Coeff. of $\frac{t^2}{2!}$ in $M_X(t) = 3\theta^2$

Hence $\quad \text{Var}(X) = \mu_2' - \mu_1'^2 = 3\theta^2 - \theta^2 = 2\theta^2$.

Example 7·3. *If the moments of variate X are defined by* $E(X^r) = 0{\cdot}6$; $r = 1, 2, 3, \ldots$ *show that* $\quad P(X=0) = 0{\cdot}4,\ P(X=1) = 0{\cdot}6,\ P(X \geq 2) = 0.$

Solution. The *m.g.f.* of variate X is :

$$M_X(t) = \sum_{r=0}^{\infty}\frac{t^r}{r!}\mu_r' = 1 + \sum_{r=1}^{\infty}\frac{t^r}{r!}(0{\cdot}6) = 0{\cdot}4 + 0{\cdot}6\sum_{r=0}^{\infty}\frac{t^r}{r!} = 0{\cdot}4 + 0{\cdot}6\,e^t \qquad \ldots(*)$$

But $$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx}\,P(X=x)$$

$$= P(X=0) + e^t\cdot P(X=1) + \sum_{x=2}^{\infty} e^{tx}\cdot P(X=x) \qquad \ldots(**)$$

From (*) and (**), we get : $\quad P(X=0) = 0{\cdot}4\ ;\ P(X=1) = 0{\cdot}6\ ;\ P(X \geq 2) = 0.$

Example 7·4. *Find the moment generating function of the random variable whose moments are :* $\quad \mu_r' = (r+1)!\,2^r.$

Solution. The *m.g.f.* is given by :

$$M_X(t) = \sum_{r=0}^{\infty}\frac{t^r}{r!}\,\mu_r' = \sum_{r=0}^{\infty}\frac{t^r}{r!}\,(r+1)!\,2^r = \sum_{r=0}^{\infty}(r+1)(2t)^r$$

$$= 1 + 2(2t) + 3(2t)^2 + 4(2t)^3 + \ldots = (1-2t)^{-2}.$$

Aliter. The R.H.S. is an arithmetic-geometric series with ratio $r = (2t)$.

Let $\quad S = 1 + 2r + 3r^2 + 4r^3 + \ldots$

Then $\quad rS = r + 2r^2 + 3r^3 + \ldots$

$$\therefore\quad (1-r)S = 1 + r + r^2 + \ldots = \frac{1}{(1-r)} \quad\Rightarrow\quad S = \frac{1}{(1-r)^2} = (1-2t)^{-2}$$

Remark. This is *m.g.f.* of Chi-square (χ^2) distribution with parameter (degrees of freedom) = 2 [*c.f.* Chapter 15.].

Example 7·5. *If* μ_r' *is the rth moment about origin, prove that*

$$\mu_r' = \sum_{j=1}^{r}\binom{r-1}{j-1}\mu'_{r-j}\,k_j, \text{ where } k_j \text{ is the } j\text{th cumulant.}$$

Solution. Differentiating both sides of (7·9*a*) in page 7·5, *w.r.to.* t, we get

$$k_1 + k_2 t + k_3\frac{t^2}{2!} + \ldots + k_r\frac{t^{r-1}}{(r-1)!} + \ldots = \frac{\mu_1' + \mu_2' t + \mu_3'\frac{t^2}{2!} + \ldots + \mu_r'\frac{t^{r-1}}{(r-1)!} + \ldots}{1 + \mu_1' t + \mu_2'\frac{t^2}{2!} + \ldots + \mu_r'\frac{t^r}{r!} + \ldots}$$

$$\Rightarrow \left\{k_1 + k_2 t + k_3\frac{t^2}{2!} + \ldots + k_r\frac{t^{r-1}}{(r-1)!} + \ldots\right\} \times \left\{1 + \mu_1' t + \mu_2'\frac{t^2}{2!} + \ldots + \mu_r'\frac{t^r}{r!} + \ldots\right\}$$

$$= \mu_1' + \mu_2' t + \mu_3'\frac{t^2}{2!} + \ldots + \mu_r'\frac{t^{r-1}}{(r-1)!} + \ldots$$

Comparing the coefficient of $\frac{t^{r-1}}{(r-1)!}$ on both sides, we get

$$\mu_r' = k_1\cdot\mu'_{r-1} + (r-1)k_2\cdot\mu'_{r-2} + \binom{r-1}{2}k_3\cdot\mu'_{r-3} + \ldots + k_r$$

$$= \binom{r-1}{0}\mu'_{r-1}k_1 + \binom{r-1}{1}\mu'_{r-2}k_2 + \binom{r-1}{2}\mu'_{r-3}k_3 + \ldots + \binom{r-1}{r-1}k_r$$

$$= \sum_{j=1}^{r}\binom{r-1}{j-1}\mu'_{r-j}k_j, \text{ which is the required result.}$$

7·3. CHARACTERISTIC FUNCTION

In some cases *m.g.f.* does not exist, since the integral $\int_{-\infty}^{\infty} e^{tx} f(x)\,dx$ or the $\sum_x e^{tx} p(x)$ does not converge absolutely for real values of t for some distributions. For example, for the continuous probability distribution :

$$dF(x) = C\frac{1}{(1+x^2)^m}dx\ ;\ m > 1,\ -\infty < x < \infty$$

the *m.g.f.* does not exist, since the integral : $M_X(t) = C\int_{-\infty}^{\infty} e^{tx}\frac{1}{(1+x^2)^m}\,dx$ does converge absolutely for finite positive values of m because the function e^{tx} does not dominate the function x^{2m} so that $e^{ax}/x^{2m} \to \infty$ as $x \to \infty$.

Again, for the discrete probability distribution :

$$f(x) = \begin{cases}\dfrac{6}{\pi^2 x^2}\ ;\ x = 1, 2, 3, \ldots\\ 0, \text{ elsewhere}\end{cases}$$

$$M_X(t) = \sum_x e^{tx}\,p(x) = \frac{6}{\pi^2}\sum_{x=1}^{\infty}\left(\frac{e^{tx}}{x^2}\right)$$

The series is not convergent (by D' Alembert's Ratio Test) for $t > 0$. Thus does not exist a positive number h such that $M_X(t)$ exist for $-h < t < h$. Hence does not exist in this case also.

A more serviceable function than the *m.g.f.* is what is known as characteristic function and is defined as

$$\phi_X(t) = E(e^{itX}) = \begin{cases}\int e^{itx} f(x)\,dx & \text{(for continuous probability distributions)}\\ \sum_x e^{itx} p(x) & \text{(for discrete probability distributions)}\end{cases}$$

If $F_X(x)$ is the distribution function of a continuous random variable X, then

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx}\,dF(x)$$

Obviously $\phi(t)$ is a complex valued function of real variable t. It may be noted that

$$|\phi(t)| = \left|\int_{-\infty}^{\infty} e^{itx} f(x)\,dx\right| \le \int_{-\infty}^{\infty} |e^{itx}|\, f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad \ldots (7\cdot11b)$$

$[\because |e^{itx}| = |\cos tx + i \sin tx|^{1/2} = (\cos^2 tx + \sin^2 tx)^{1/2} = 1]$

Since $|\phi(t)| \le 1$, characteristic function $\phi_X(t)$ always exists.

Remarks 1.
$$\phi_X(t) = E(e^{itX}) = E\left[1 + itX + \frac{(it)^2}{2!}X^2 + \ldots + \frac{(it)^r}{r!}X^r + \ldots\right]$$
$$= 1 + it\,E(X) + \frac{(it)^2}{2!}E(X^2) + \ldots + \frac{(it)^r}{r!}E(X^r) + \ldots$$
$$= 1 + it\,\mu_1' + \frac{(it)^2}{2!}\mu_2' + \ldots + \frac{(it)^r}{r!}\mu_r' + \ldots \qquad \ldots (7\cdot11c)$$

where $\mu_r' = EX^r$, is the rth moment of X about origin.

$\therefore$ μ_r' = Coefficient of $\frac{(it)^r}{r!}$ in $\phi_X(t)$ ... (7·11d)

Hence, like *m.g.f.*, the characteristic function $\phi(t)$, also generates moments.

Re-writing (7·11 *b*), we get

$$\phi_X(t) = \sum_{r=0}^{\infty} \frac{(it)^r}{r!}\mu_r' \qquad \ldots (7\cdot11e)$$

2. Cumulant generating function $K_X(t)$ in terms of $\phi_X(t)$ is given by :

$$K(t) = \log \phi(t) \equiv \kappa_1 \cdot (it) + \kappa_2 \cdot \frac{(it)^2}{2!} + \ldots + \kappa_r \cdot \frac{(it)^r}{r!} + \ldots \qquad \ldots (7\cdot12)$$
$$= \sum_{r=1}^{\infty} \frac{(it)^r}{r!}\kappa_r,$$

here, κ_r = Coefficient of $\frac{(it)^r}{r!}$ in $K_X(t)$, is the rth cumulant.

3. Yet another advantage of characteristic function lies in the fact that it uniquely termines the distribution function, *i.e.*, if the characteristic function of a distribution is given, e distribution can be uniquely determined by the theorem, known as the *Uniqueness Theorem Characteristic Functions* discussed later in the chapter [*Theorem 7·2*, Page 7·21].

3·1. PROPERTIES OF CHARACTERISTIC FUNCTION

Property 1. For all real t, we have

(*i*) $\phi(0) = \int_{-\infty}^{\infty} dF(x) = 1$ ... (7·13*a*)

(*ii*) $|\phi(t)| \le 1 = \phi(0)$ [From (7·11 *b*)] ... (7·13*b*)

Property 2. *$\phi(t)$ is continuous everywhere, i.e., $\phi(t)$ is continuous function of t in $(-\infty, \infty)$. Rather $\phi(t)$ is uniformly continuous in 't'.*

Proof. For $h \ne 0$, $|\phi_X(t+h) - \phi_X(t)| = \left|\int_{-\infty}^{\infty} \left[e^{i(t+h)x} - e^{itx}\right] dF(x)\right|$

$$\le \int_{-\infty}^{\infty} \left|e^{itx}(e^{ihx} - 1)\right| dF(x) = \int_{-\infty}^{\infty} \left|e^{ihx} - 1\right| dF(x) \quad \ldots (*)$$

The last integral does not depend on t. If it tends to zero as $h \to 0$ then $\phi_X(t)$ is uniformly continuous in 't'.

Now $|e^{ihx} - 1| \le |e^{ihx}| + 1| \le 1 + 1 = 2$

$\therefore$ $\int_{-\infty}^{\infty} \left|e^{ihx} - 1\right| dF(x) \le 2\int_{-\infty}^{\infty} dF(x) = 2$

Hence by Dominated Convergence Theorem (D.C.T.), taking the limit inside the integral sign in (*), we get

$$\lim_{h\to 0} \left|\phi_X(t+h) - \phi_X(t)\right| \le \int_{-\infty}^{\infty} \lim_{h\to 0} \left|e^{ihx} - 1\right| dF(x) = 0$$

$\Rightarrow$ $\lim_{h\to 0} \phi_X(t+h) = \phi_X(t), \forall\, t$

Hence $\phi_X(t)$ is uniformly continuous in 't'.

Property 3. *$\phi_X(-t)$ and $\phi_X(t)$ are conjugate functions, i.e., $\phi_X(-t) = \overline{\phi_X(t)}$, where $\bar{a}$ is the complex conjugate of a.*

Proof. $\phi_X(t) = E(e^{itX}) = E[\cos tX + i \sin tX]$

$\therefore$ $\overline{\phi_X(t)} = E(\cos tX - i \sin tX) = E[\cos(-t)X + i \sin(-t)X] = E(e^{-itX}) =$

Property 4. *If the distribution function of a r.v. X is symmetrical about zero, i.e.,*

$$1 - F(x) = F(-x) \quad \Rightarrow \quad f(-x) = f(x),$$

then $\phi_X(t)$ is real valued and even function of t.

Proof. By *def.*, we have

$\therefore$ $\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx = \int_{-\infty}^{\infty} e^{-ity} f(-y)\,dy$

$= \int_{-\infty}^{\infty} e^{-ity} f(y)\,dy$ $[\because$

$= \phi_X(-t)$

$\Rightarrow$ $\phi_X(t)$ is an even function of t.

From (7·13c) and (7·13d), we get : $\phi_X(t) = \phi_X(-t) = \overline{\phi_X(t)}$

Hence $\phi_X(t)$ is a real valued and even function of t.

Property 5. *If X is some r.v. with characteristic function $\phi_X(t)$, and exists,*

then $\mu_r' = (-i)^r \left|\frac{\partial^r}{\partial t^r}\phi(t)\right|_{t=0}$

Proof. $\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx$

Differentiating (under the integral sign) 'r' times w.r. to t, we get

$$\frac{\partial^r}{\partial t^r}\phi(t) = \int_{-\infty}^{\infty} (ix)^r e^{itx} f(x)\,dx = (i)^r \int_{-\infty}^{\infty} x^r e^{itx} f(x)\,dx$$

$$\therefore \left|\frac{\partial^r}{\partial t^r}\phi(t)\right|_{t=0} = (i)^r \left|\int_{-\infty}^{\infty} x^r e^{itx} f(x)\,dx\right|_{t=0}$$

$$= (i)^r \int_{-\infty}^{\infty} x^r f(x)\,dx = i^r E(X^r) = i^r \mu_r'$$

Hence $$\mu_r' = \left(\frac{1}{i}\right)^r \left|\frac{\partial^r}{\partial t^r}\phi(t)\right|_{t=0} = (-i)^r \left|\frac{\partial^r}{\partial t^r}\phi(t)\right|_{t=0}$$

Alter. We have : [*c.f.* equation (7·11*b*)]

$$\phi_X(t) = 1 + it\mu_1' + \frac{(it)^2}{2!}\mu_2' + \ldots + \frac{(it)^r}{r!}\mu_r' + \frac{(it)^{r+1}}{(r+1)!}\mu_{r+1}' + \ldots$$

Differentiating r times w.r. to t and then putting $t = 0$, we get

$$\left|\frac{\partial^r}{\partial t^r}\phi_X(t)\right|_{t=0} = i^r \mu_r' \quad \Rightarrow \quad \mu_r' = (-i)^r \left|\frac{\partial^r}{\partial t^r}\phi_X(t)\right|_{t=0} \quad \left(\because i = -\frac{1}{i}\right)$$

Property 6. *$\phi_{cX}(t) = \phi_X(ct)$, c being a constant.*

Property 7. *If X_1 and X_2 are independent random variables, then*

$$\phi_{X_1 + X_2}(t) = \phi_{X_1}(t)\,\phi_{X_2}(t) \qquad \ldots (7\cdot14a)$$

More generally, for independent random variables X_i; $i = 1, 2, .., n$, we have

$$\phi_{X_1 + X_2 + \ldots + X_n}(t) = \phi_{X_1}(t)\,\phi_{X_2}(t) \ldots \phi_{X_n}(t) \qquad (7\cdot14b)$$

Important Remarks. Converse of (*) is not true, *i.e.*,

$\phi_{X_1 + X_2}(t) = \phi_{X_1}(t)\,\phi_{X_2}(t)$, does not imply that X_1 and X_2 are independent.

For example, let X_1 be a standard Cauchy variate with *p.d.f.* :

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

Then $\phi_{X_1}(t) = e^{-|t|}$ (*c.f.* Chapter 9) (*)

Let $X_2 = X_1$, almost surely (*a.s.*), *i.e.*, $P(X_1 = X_2) = 1$. (**)

Then $\phi_{X_2}(t) = e^{-|t|}$

Now $\phi_{X_1 + X_2}(t) = \phi_{2X_1}(t) = \phi_{X_1}(2t) = e^{-2|t|} = \phi_{X_1}(t)\,\phi_{X_2}(t)$ [From (**) and (*)]

i.e., (7·14*a*) is satisfied but obviously X_1 and X_2 are not independent, because of (**).

In fact, (7·14*a*) will hold even if we take $X_1 = aX$ and $X_2 = bX$, a and b being real numbers so that X_1 and X_2 are connected by the relation : $\frac{X_1}{a} = X = \frac{X_2}{b} \Rightarrow aX_2 = bX_1$.

As another example let us consider the joint *p.d.f.* of two r.v.'s X and Y given by :

$$f(x, y) = \begin{cases} \frac{1}{4a^2}\left[1 + xy\left(x^2 - y^2\right)\right]; & |x| \le a,\ |y| \le a,\ a > 0 \\ 0, \text{ elswhere} \end{cases}$$

Then the marginal *p.d.f.*'s of X and Y are given by :

$$g(x) = \int_{-a}^{a} f(x, y)\,dy = \frac{1}{2a}; \quad |x| \le a$$

$$h(y) = \int_{-a}^{a} f(x, y)\,dx = \frac{1}{2a}; \quad |y| \le a$$

Then $$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} g(x)\,dx = \frac{1}{2a}\int_{-a}^{a} e^{itx}\,dx = \frac{e^{ait} - e^{-ait}}{2ait} = \frac{\sin at}{at}$$

Similarly, $$\phi_Y(t) = \frac{\sin at}{at} \qquad \therefore \phi_X(t)\,\phi_Y(t) = \left(\frac{\sin at}{at}\right)^2$$

The *p.d.f.* $k(z)$ of the r.v. $Z = X + Y$ is given by the convolution of *p.d.f.*'s of X and Y, *viz.*,

$$k(z) = \int f(u, z - u)\,du = \frac{1}{4a^2}\int \left[1 + u(z - u)\left[u^2 - (z - u)^2\right]\right] du$$

$$= \frac{1}{4a^2}\int (1 + 3z^2u^2 - 2zu^3 - z^3u)\,du,$$

the limits of integration for u being in terms of z and are given by :

$$-a \le u \le z + a;\ u \le 0 \quad \text{and} \quad z - a \le u \le a;\ u > 0$$

(Left as an exercise to the reader.)

Thus

$$k(z) = \begin{cases} \frac{1}{4a^2}\int_{-a}^{z+a} (1 + 3z^2u^2 - 2zu^3 - z^3u)\,du = \frac{2a + z}{4a^2}; & -2a \le z \le 0 \\ \frac{1}{4a^2}\int_{z-a}^{a} (1 + 3z^2u^2 - 2zu^3 - z^3u)\,du = \frac{2a - z}{4a^2}; & 0 < z \le 2a \\ 0, \text{ elsewhere} \end{cases}$$

Now

$$\phi_{X+Y}(t) = \phi_Z(t) = \int_{-2a}^{2a} e^{itz}\,k(z)\,dz$$

$$= \int_{-2a}^{0} \left(\frac{2a + z}{4a^2}\right) e^{itz}\,dz + \int_{0}^{2a} \left(\frac{2a - z}{4a^2}\right) e^{itz}\,dz$$

$$= \int_{0}^{2a} \left(e^{-itz} + e^{itz}\right)\left(\frac{2a - z}{4a^2}\right) dz \qquad \text{[Changing } z \text{ to } -z \text{ in the first integ}$$

$$= \frac{1}{2a^2}\int_{0}^{2a} (2a - z)\cos tz\,dz = \frac{2 - 2\cos 2at}{4a^2t^2} = \frac{1 - \cos 2at}{2a^2t^2} = \left(\frac{\sin at}{at}\right)^2$$

$$= \phi_X(t) \cdot \phi_Y(t)$$

But $g(x).\,h(y) \ne f(x, y) \Rightarrow$ X and Y are not independent.

However, the joint characteristic function:

$$\phi_{X_1, X_2}(t_1, t_2) = E\left(e^{i(t_1X_1 + t_2X_2)}\right) = \phi_{X_1}(t) \cdot \phi_{X_2}(t),$$

$\Rightarrow$ X_1 and X_2 are independent. [For details, see *Theorem* 7·3, page 7·21]

Property 8. *Effect of Change of Origin and Scale on Characteristic Function.*

$U = \frac{X - a}{h}$, a and h being constants, then $\phi_U(t) = e^{-iat/h}\,\phi_X(t/h)$.

In particular if we take $a = E(X) = \mu$ (say) and $h = \sigma_X = \sigma$, then the characteristic function of the standard variate $Z = \frac{X - E(X)}{\sigma_X} = \frac{X - \mu}{\sigma}$, is given by :

$$\phi_Z(t) = e^{-i\mu t/\sigma}\,\phi_X(t/\sigma)$$

Proofs of properties 6, 7 and 8 are similar to the properties 1, 2 and 3 of moment erating functions.

Definition. *A random variable X is said to be a Lattice variable or be lattice distributed, if some $h > 0$, $P[(X/h) \text{ is an integer}] = 1$; h is called a mesh.*

Property 9. *If $|\phi_X(s)| = 1$ for some $s \neq 0$, then for some real a, $X - a$ is a lattice variable mesh, $h = 2\pi/|s|$.*

Proof. Consider any fixed t. We can write

$$\phi_X(t) = |\phi_X(t)| e^{i\omega t}, \text{ (a dependent on t),}$$

e any complex number z can be written as $z = |z| e^{i\theta}$.

$$\therefore \quad |\phi_X(t)| = e^{-ia t}\phi_X(t) = \phi_{X-a}(t)$$

$$= E\{\cos t(X-a) + i \sin t(X-a)\} = E\{\cos t(X-a)\}$$

e left-hand side being real, we must have $E\{\sin t(X-a)\} = 0$.

$$\therefore \quad 1 - |\phi_X(t)| = E\{1 - \cos t(X-a)\} \qquad \ldots (*)$$

If $|\phi_X(s)| = 1$, for some $s \neq 0$, then for some a dependent on s, we have from (*)

$$E\{1 - \cos s(X-a)\} = 0 \qquad \ldots (**)$$

But since $1 - \cos s(X-a)$ is a non-negative random variable, (**) implies that

$$P\{1 - \cos s(X-a) = 0\} = 1 \quad \Rightarrow \quad P[\cos s(X-a) = 1] = 1 \quad \Rightarrow \quad P[s(X-a) = 2n\pi] = 1$$

$$\therefore \quad P\left[(X-a) = \frac{2n\pi}{|s|}\right] = 1, \text{ for some } n = 0, 1, 2, \ldots$$

Thus $(X-a)$ is a Lattice variable with mesh $h = \dfrac{2\pi}{|s|}$.

7·3·2. Necessary and Sufficient Conditions for a Function $\phi(t)$ to be aracteristic Function. Properties 1 to 3 in § 7·3·1 are merely the necessary ditions for a function $\phi(t)$ to be the characteristic function of a *r.v.* X. Thus, if a ction $\phi(t)$ does not satisfy any one of these conditions, it cannot be the characteristic ction of any *r.v.* X, *e.g.* the function $\phi(t) = \log(1+t)$ cannot be the *c.f.* of *r.v.* X,

ce $\phi(0) = \log 1 = 0 \neq 1$.

These conditions are, however, not sufficient. It has been shown (*c.f.* Methods of thematical Statistics by H. Cramer) that if $\phi(t)$ is near $t = 0$ of the form,

$$\phi(t) = 1 + o(t^2 + \delta), \ \delta > 0 \qquad \ldots (*)$$

here $O(t^r)$ divided by t^r tends to zero as $t \to 0$, then $\phi(t)$ cannot be the characteristic nction unless it is identically equal to one. Thus, the functions

$$(i)\ \phi(t) = e^{-t^4} = 1 + o(t^4), \quad \text{and} \quad (ii)\ \phi(t) = \frac{1}{1+t^4} = 1 + o(t^4),$$

ng of the form (*) are not characteristic functions, though both satisfy all the cessary conditions.

We give below a set of sufficient but not necessary conditions, due to Polya, for a nction $\phi(t)$ to be the characteristic function :

$\phi(t)$ is a characteristic function if

(1) $\phi(0) = 1$, (2) $\phi(t) = \phi(-t)$, (3) $\phi(t)$ is continuous,

(4) $\phi(t)$ is convex for $t > 0$, *i.e.*, for $t_1, t_2 > 0$, $2\phi\left[\frac{1}{2}(t_1 + t_2)\right] \leq \phi(t_1) + \phi(t_2)$

(5) $\lim_{t \to \infty} \phi(t) = 0$.

Hence by Polya's conditions, the functions $e^{-|t|}$ and $[1 + |t|]^{-1}$ are characteristic functions. However, Polya's conditions are only sufficient and not necessary for a characteristic function.

For example if $X \sim N(\mu, \sigma^2)$, *i.e.*, X has normal distribution with mean μ and variance σ^2, then :

$$\phi(t) = \exp\{it\mu - (t^2\sigma^2)/2\} \text{ and } \phi(-t) \neq \phi(t).$$

Various necessary and sufficient conditions are known, the simplest seems to be the following, due to Cramer.

"In order that a given, bounded and continuous function $\phi(t)$ should be the characteristic function of a distribution, it is necessary and sufficient that $\phi(0) = 1$ and that the function

$$\phi(x, A) = \int_0^A \int_0^A \phi(t-u) \exp[(t-u)ix]\, dt\, du$$

is real and non-negative for all x and all $A > 0$.

7·3·3. Multivariate Characteristic Function. Let

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \text{and} \quad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, \ t \text{ real}$$

be $n \times 1$ column vectors. Then characteristic function of X is defined as :

$$\phi_X(t) = E(e^{it'X}) = E[e^{i(t_1X_1 + t_2X_2 + \ldots + t_nX_n)}] \qquad \ldots (7$$

We may also write it as $\phi_{X_1, X_2, \ldots, X_n}(t_1, t_2, \ldots, t_n)$ or $\phi_X(t_1, t_2, \ldots, t_n)$

Some Properties

(i) $\phi_X(0, 0, \ldots, 0) = 1$, (ii) $\phi_{-X}(t) = \overline{\phi_X(t)}$, (iii) $|\phi_X(t)| \leq 1$,

(iv) $\phi_X(t)$ is uniformly continuous in n-dimensional Euclidian space,

(v) If $f_X(\cdot)$ is *p.d.f.* of X,

$$\phi_X(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, x_2, \ldots, x_n) \exp\{i \sum_j t_j x_j\} dx_1\, dx_2 \ldots dx_n$$

(vi) $\phi_X(t) = \phi_Y(t)$ for all t, then X and Y have the same distribution.

(vii) If $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\phi_X(t_1, t_2, \ldots, t_n)|\, dt_1\, dt_2 \ldots dt_n < \infty$, then X is abso continuous and has a uniformly continuous *p.d.f.*

$$f_X(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(i \sum_j t_j x_j)\ \phi_X(t_1, t_2, \ldots, t_n)\, dt_1$$

(viii) The random variables $X_1, X_2, \ldots, X_n$ are (mutually) independent iff

$$\phi_{X_1, X_2, \ldots, X_n}(t_1, t_2, \ldots, t_n) = \phi_{X_1}(t_1)\, \phi_{X_2}(t_2) \ldots \phi_{X_n}(t_n)$$

TABLE 9·2. AREAS UNDER NORMAL CURVE

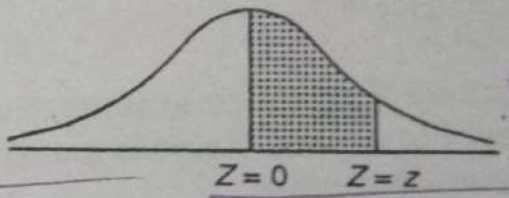
Normal probability curve is given by :

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

and standard normal probability curve is given by :

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), \quad -\infty < z < \infty$$

where $Z = \frac{X - E(X)}{\sigma_x} \sim N(0, 1)$



The following table gives the shaded area in the diagram, *viz.*, $P(0 < Z < z)$ for different values of z.

TABLE OF AREAS

↓Z →	0	1	2	3	4	5	6	7	8	9
·0	·0000	·0040	·0080	·0120	·0160	·0199	·0239	·0279	·0319	·0359
·1	·0398	·0438	·0478	·0517	·0557	·0596	·0636	·0675	·0714	·0759
·2	·0793	·0832	·0871	·0910	·0948	·0987	·1026	·1064	·1103	·1141
·3	·1179	·1217	·1255	·1293	·1331	·1368	·1406	·1443	·1480	·1517
·4	·1554	·1591	·1628	·1664	·1700	·1736	·1772	·1808	·1844	·1879
·5	·1915	·1950	·1985	·2019	·2054	·2088	·2123	·2157	·2190	·2224
·6	·2257	·2291	·2324	·2357	·2389	·2422	·2454	·2486	·2517	·2549
·7	·2580	·2611	·2642	·2673	·2703	·2734	·2764	·2794	·2823	·2852
·8	·2881	·2910	·2939	·2967	·2995	·3023	·3051	·3078	·3106	·3133
·9	·3159	·3186	·3212	·3238	·3264	·3289	·3315	·3340	·3365	·3389
1·0	·3413	·3438	·3461	·3485	·3508	·3531	·3554	·3577	·3599	·3621
1·1	·3643	·3655	·3686	·3708	·3729	·3749	·3770	·3790	·3810	·3830
1·2	·3849	·3869	·3888	·3907	·3925	·3944	·3962	·3980	·3997	·4015
1·3	·4032	·4049	·4066	·4082	·4099	·4115	·4131	·4147	·4162	·4177

10.2 INTRODUCTION

So far we have confined ourselves to univariate distributions, *i.e.*, the ributions involving only one variable. Often we come across situations in which focus is simultaneously on two or more variables and invariably, we observe that vements in one variable are accompanied by movements in other variable. For mple, husband's age and wife's age move together, scores on an I.Q. test move h scores in university examinations. Similarly, studies in income and expenditure households or price and demand of commodities, exhibit accompaning movements wo variables. Notwithstanding the cases of spurious or nonsensical relations ch we may enjoy through such funny combinations of variables as suggesting that number of runs scored by a batsman increases with an increase in the sumption of fertilizer in the local market or the number of flights space is easing with a decrease in the population of tigers, the study of variables indicating mpaning behaviour is of great interest in Statistics.

However, the above statements are not precise enough to be of use to decision kers. We are therefore on the look out for a *quantitative measure* of the relationship ween the two variables, and also for an appropriate mathematical or statistical of the relationship. While the second question will be discussed in the next ter, we will take up the first question in the present chapter.

10.3 MEANING OF CORRELATION

In a bivariate distribution we may be interested to find out if there is any lation or covariation between the two variables under study. If the change in one able affects a change in the other variable, the variables are said to be correlated. If wo variables deviate in the same direction, *i.e.*, if the increase (or decrease) in one ts in a corresponding increase (or decrease) in the other, correlation is said to be or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if ase (or decrease) in one results in corresponding decrease (or increase) in the correlation is said to be *diverse* or *negative*. For example, the correlation between heights and weights of a group of persons, and (*ii*) the income and expenditure; itive and the correlation between (*i*) price and demand of a commodity and (*ii*) olume and pressure of a perfect gas; is negative. Correlation is said to be *perfect* if eviation in one variable is followed by a corresponding and proportional tion in the other.

10.3.1 SCATTER DIAGRAM

is the simplest way of the diagrammatic representation of bivariate data. Thus e bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$, if the values of the variables X and Y lotted along the x-axis and y-axis respectively in the x-y plane, the diagram of o obtained is known as *scatter diagram*. From the scatter diagram, we can form a good, though vague, idea whether the variables are correlated or not, *e.g.*, if the are very dense, *i.e.*, very close to each other, we should expect a fairly good nt of correlation between the variables and if the points are widely scattered, a correlation is expected. This method, however, is not suitable if the number of vations is fairly large.

10.4 KARL PEARSON'S COEFFICIENT OF CORRELATION

As a measure of itensity or degree of linear relationship between two variables, Karl Pearson (1867-1936), *a British Biometrician*, developed a formula called *Correlation Coefficient*.

Correlation coefficient between two random variables X and Y, usually denoted by $r(X, Y)$ or simply r_{XY}, is a numerical measure of *linear relationship* between them and is defined as :

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad \ldots (10\text{-}1)$$

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

$$\left.\begin{aligned} \text{Cov}(X, Y) &= E[\{X - E(X)\}\{Y - E(Y)\}] \\ &= \frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y}) = \mu_{11} \\ \sigma_X^2 &= E\{X - E(X)\}^2 = \frac{1}{n}\Sigma (x_i - \bar{x})^2 \\ \sigma_Y^2 &= E\{Y - E(Y)\}^2 = \frac{1}{n}\Sigma (y_i - \bar{y})^2 \end{aligned}\right\} \quad \ldots (10\text{-}2)$$

the summation extending over i from 1 to n.

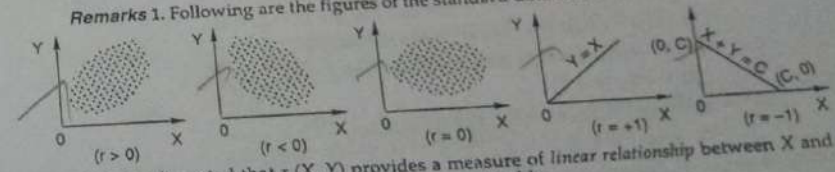
Another convenient form of the formula (10-2) for computational work is as follows :

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

$$= \frac{1}{n}\Sigma x_i y_i - \bar{y}\frac{1}{n}\Sigma x_i - \bar{x}\frac{1}{n}\Sigma y_i + \bar{x}\bar{y}$$

$$\therefore \quad \text{Cov}(X, Y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}, \quad \sigma_X^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 \text{ and } \sigma_Y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 \quad \ldots (10\text{-}2a)$$

Remarks 1. Following are the figures of the standard data for $r > 0, < 0, = 0$, and $r = \pm 1$:



2. It may be noted that $r(X, Y)$ provides a measure of *linear relationship* between X and Y. For non-linear relationship, however, it is not very suitable.

3. Sometimes, we write : $\text{Cov}(X, Y) = \sigma_{XY}$.

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient*, since

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

10.4.1. Limits for Correlation Coefficient. We have

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma (x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma (y_i - \bar{y})^2\right]^{1/2}}$$

$$\therefore \quad r^2(X, Y) = \frac{\left(\sum_i a_i b_i\right)^2}{\left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right)} \quad \text{where} \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix}$$

We have the Schwartz inequality which states that if a_i, b_i, $i = 1, 2, \ldots, n$ are real quantities then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right),$$

the sign of equality holding if and only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$

Using Schwartz inequality, we get from (*) :

$$r^2(X, Y) \le 1 \text{ i.e., } |r(X, Y)| \le 1 \quad \Rightarrow \quad -1 \le r(X, Y) \le 1 \qquad \ldots(10\cdot3)$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between -1 and $+1$. *If $r = +1$, the correlation is perfect and positive and if $r = -1$, correlation is perfect and negative.*

Aliter. If we write $E(X) = \mu_X$ and $E(Y) = \mu_Y$, then $E\left[\left(\frac{X-\mu_X}{\sigma_X}\right) \pm \left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]^2 \ge 0$

$$\Rightarrow E\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \ge 0 \quad \Rightarrow \quad 1 + 1 \pm 2\, r(X, Y) \ge 0$$

$$\therefore \quad -1 \le r(X, Y) \le 1.$$

Theorem 10·1. *Correlation coefficient is independent of change of origin and scale.*

Proof. Let $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, so that $X = a + hU$ and $Y = b + kV$, where a, b, h, k are constants; $h > 0$, $k > 0$.

We shall prove that $r(X, Y) = r(U, V)$.

Since $X = a + hU$ and $Y = b + kV$, on taking expectations, we get

$$E(X) = a + hE(U) \quad \text{and} \quad E(Y) = b + kE(V)$$

$$\therefore \quad X - E(X) = h[U - E(U)] \quad \text{and} \quad Y - E(Y) = k[V - E(V)]$$

$$\Rightarrow \quad \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E[h\{U - E(U)\}\, k\{V - E(V)\}]$$

$$= hk\, E[\{U - E(U)\}\{V - E(V)\}] = hk\, \text{Cov}(U, V) \qquad \ldots(10\cdot4)$$

$$\sigma_X^2 = E[\{X - E(X)\}^2] = E[h^2\{U - E(U)\}^2] = h^2\sigma_U^2$$

$$\Rightarrow \quad \sigma_X = h\sigma_U, (h > 0) \qquad \ldots(10\cdot4a)$$

and $$\sigma_Y^2 = E[\{Y - E(Y)\}^2] = E[k^2\{V - E(V)\}^2] = k^2\sigma_V^2$$

$$\Rightarrow \quad \sigma_Y = k\sigma_V, (k > 0) \qquad \ldots(10\cdot4b)$$

Substituting from (10·4), (10·4*a*) and (10·4*b*) in (10·1), we get

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{hk\, \text{Cov}(U, V)}{hk\, \sigma_U \sigma_V} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = r(U, V).$$

Remark. This theorem is of fundamental importance in the numerical computation of the correlation coefficient.

Corollary. *If X and Y are random variables and a, b, c, d are any numbers provided only that $a \ne 0$, $c \ne 0$, then*

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y). \qquad \ldots(10\cdot4c)$$

Proof. With usual notations, we have

$$\text{Var}(aX + b) = a^2\sigma_X^2; \qquad \text{Var}(cY + d) = c^2\sigma_Y^2;$$

$$\text{Cov}(aX + b, cY + d) = ac\sigma_{XY}$$

$$\therefore \quad r(aX + b, cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{[\text{Var}(aX + b)\,\text{Var}(cY + d)]^{1/2}} = \frac{ac\,\sigma_{XY}}{|a|\,|c|\,\sigma_X \sigma_Y} = \frac{ac}{|ac|} r(X, Y)$$

If $ac > 0$, *i.e.*, if a and c are of same signs, then $ac/|ac| = +1$.

If $ac < 0$, *i.e.*, if a and c are of opposite signs, then $ac/|ac| = -1$.

Remark. In particular, if we take $b = 0 = d$, then from (10·4*c*), we get

$$r(aX, cY) = \frac{ac}{|ac|} r(X, Y).$$

Theorem 10·2. *Two independent variables are uncorrelated.*

Proof. If X and Y are independent variables, then

$$\text{Cov}(X, Y) = 0 \quad \Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the *converse* of the theorem is not true, *i.e.*, two uncorrelated variables may not be independent as the following example illustrates :

X	−3	−2	−1	1	2	3	Total
							$\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	−27	−8	−1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0, \qquad \text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = 0$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Thus in the above example, the variables X and Y are uncorrelated. But on careful examination we find that X and Y are not independent but they are connected by the relation $Y = X^2$. Hence two uncorrelated variables need not necessarily be independent. A simple reasoning for this strange conclusion is that $r(X, Y) = 0$, merely implies the absence of any linear relationship between the variables X and Y. There may, however, exist some other form of relationship between them, e.g., quadratic, cubic or trigonometric.

Remarks **1.** Following are some more examples where two variables are uncorrelated but *not* independent.

(*i*) $X \sim N(0, 1)$ and $Y = X^2$. Since $X \sim N(0, 1)$, $E(X) = 0 = E(X^3)$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = E(X^3) - E(X)\,E(Y) = 0$$

$$\Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0.$$

Hence X and Y are uncorrelated but not independent ($\because Y = X^2$).

(*ii*) Let X be a r.v. with p.d.f. $f(x) = \frac{1}{2}$, $-1 \le x \le 1$, and let $Y = X^2$.

Here we shall get

$$E(X) = 0 \text{ and } E(XY) = E(X^3) = 0, \text{ so that } r(X, Y) = 0$$

2. However, the converse of the theorem holds in the following cases :

(*a*) If X and Y are jointly normally distributed with $\rho = \rho(X, Y) = 0$, then they are independent. If $\rho = 0$, then [*c.f.* § 11·3, Equation (11·7)]

$= \frac{ac}{|ac|} r(X, Y)$

σ_Y

... (10·4d)

ed variables may

$= 0$

$= 28$

$= 0$

$= 0$

But on careful
nected by the
cessarily be
$) = 0$, merely
nd Y. There
, quadratic,

orrelated but

$(\because Y = X^2)$

n they are

$$f(x, y) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu_x}{\sigma_x}\right)^2\right] \times \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y-\mu_y}{\sigma_y}\right)^2\right] = f_1(x)\, f_2(y)$$

Hence X and Y are independent.

(b) If each of the two variables X and Y takes two values, 0, 1 with positive probabilities, then $r(X, Y) = 0 \Rightarrow X$ and Y are independent.

Proof. Let X take the values 1 and 0 with positive probabilities p_1 and q_1 respectively and take the values 1 and 0 with positive probabilities p_2 and q_2 respectively. Then

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}(X, Y) = 0$$

$$\Rightarrow \quad 0 = E(XY) - E(X)\,E(Y) = 1 \,.\, P(X = 1 \cap Y = 1) - [1 \,.\, P(X = 1) \times 1 \,.\, P(Y = 1)]$$

$$= P(X = 1 \cap Y = 1) - p_1p_2$$

$$\Rightarrow \quad P(X = 1 \cap Y = 1) = p_1p_2 \quad = P(X = 1) \,.\, P(Y = 1). \text{ Hence } X \text{ and } Y \text{ are independent.}$$

10·4·2. Assumptions Underlying Karl Pearson's Correlation Coefficient. onian correlation coefficient r is based on the following assumptions :

(i) The variables X and Y under study are linearly related. In other words, the scatter am of the data will give a straight line curve.

(ii) Each of the variables (series) is being affected by a large number of independent ibutory causes of such a nature as to produce normal distribution. For example, the bles (series) relating to ages, heights, weights, supply, price, etc., conform to this mption. In the words of Karl Pearson :

"The sizes of the complex of organs (something measurable) are determined by a great y of independent contributory causes, for example, climate, nourishment, physical g and innumerable other causes which cannot be individually observed or their effects ired." Karl Pearson further observes, "The variations in intensity of the contributory are small as compared with their absolute intensity and these variations follow the l law of distribution."

(iii) The forces so operating on each of the variable series are not independent of each other related in a causal fashion. In other words, cause and effect relationship exists en different forces operating on the items of the two variable series. These forces e common to both the series. If the operating forces are entirely independent of other and not related in any fashion, then there cannot be any correlation en the variables under study.

r example, the correlation coefficient between,

the series of heights and incomes of individuals over a period of time,

the series of marriage rate and the rate of agricultural production in a country over a period of time, and

the series relating to the size of the shoe and intelligence of a group of individuals,

be zero, since the forces affecting the two variable series in each of the above re entirely independent of each other.

owever, if in any of the above cases the value of r for a given set of data is not hen such correlation is termed as *chance correlation* or *spurious* or *nonsense* tion.

Example 10·7. *Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :*

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

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

$$\overline{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68, \qquad \overline{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \overline{X}\,\overline{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \overline{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \overline{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left\{\frac{37028}{8} - (68)^2\right\}\left\{\frac{38132}{8} - (69)^2\right\}}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628{\cdot}5 - 4624)\,(4766{\cdot}5 - 4761)}} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603.$$

Aliter. (SHORT-CUT METHOD)

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2
65	67	−3	−2	9	4
66	68	−2	−1	4	1
67	65	−1	−4	1	16
67	68	−1	−1	1	1
68	72	0	3	0	9
69	72	1	3	1	9
70	69	2	0	4	0
72	71	4	2	16	4
Total		0	0	36	44

$$\overline{U} = \frac{1}{n}\Sigma U = 0, \quad \overline{V} = \frac{1}{n}\Sigma V = 0, \quad \text{Cov}(U, V) = \frac{1}{n}\Sigma UV - \overline{U}\,\overline{V} = \frac{1}{8} \times 24 = 3$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - (\overline{U})^2 = \frac{1}{8} \times 36 = 4.5, \quad \sigma_V^2 = \frac{1}{n}\Sigma V^2 - (\overline{V})^2 = \frac{1}{8} \times 44 = 5.5$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4.5 \times 5.5}} = 0.603 = r(X, Y)$$

Remark. The reader is advised to calculate the correlation coefficient by arbitrary origin method rather than by the direct method, since the latter leads to much simpler arithmetical calculations.

Example 10.2. *A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$n = 25, \quad \Sigma X = 125, \quad \Sigma X^2 = 650, \quad \Sigma Y = 100, \quad \Sigma Y^2 = 460, \quad \Sigma XY = 508$

If was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

. Obtain the correct value of correlation coefficient.

Solution.

Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$

$$\overline{X} = \frac{1}{n}\Sigma X = \frac{1}{25} \times 125 = 5, \qquad \overline{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25} \times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \overline{X}\,\overline{Y} = \frac{1}{25} \times 520 - 5 \times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \overline{X}^2 = \frac{1}{25} \times 650 - (5)^2 = 1; \qquad \sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \overline{Y}^2 = \frac{1}{25} \times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{4/5}{|1 \times (6/5)|} = \frac{2}{3} = 0.67$$

Example 10.3. *Show that if X′, Y′ are the deviations of the random variables X and Y from their respective means, then*

$$(i) \quad r = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2 \qquad (ii) \quad r = -1 + \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y}\right)^2.$$

Deduce that $-1 \le r \le +1$.

Solution. (*i*) Here $X_i' = (X_i - \overline{X})$ and $Y_i' = (Y_i - \overline{Y})$

$$\text{R.H.S.} = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2 = 1 - \frac{1}{2N}\sum_i \left[\frac{X_i'^2}{\sigma_X^2} + \frac{Y_i'^2}{\sigma_Y^2} - \frac{2X_i'Y_i'}{\sigma_X \sigma_Y}\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i X_i'^2 + \frac{1}{\sigma_Y^2}\sum_i Y_i'^2 - \frac{2}{\sigma_X \sigma_Y}\sum_i X_i'Y_i'\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i (X_i - \overline{X})^2 + \frac{1}{\sigma_Y^2}\sum_i (Y_i - \overline{Y})^2 - \frac{2}{\sigma_X \sigma_Y}\sum_i (X_i - \overline{X})(Y_i - \overline{Y})\right]$$

$$= 1 - \frac{1}{2}\left(\frac{1}{\sigma_X^2} \cdot \sigma_X^2 + \frac{1}{\sigma_Y^2} \cdot \sigma_Y^2 - \frac{2}{\sigma_X \sigma_Y} \cdot r\sigma_X \sigma_Y\right) = 1 - \frac{1}{2}(1 + 1 - 2r) = r$$

(*ii*) Proceeding similarly, we will get R.H.S. $= -1 + \frac{1}{2}(1 + 1 + 2r) = r$

Deduction. Since $\left(\frac{X_i'}{\sigma_X} \pm \frac{Y_i'}{\sigma_Y}\right)^2$, being the square of a real quantity is always non-negative, $\sum_i \left(\frac{X_i'}{\sigma_X} \pm \frac{Y_i'}{\sigma_Y}\right)^2$ is also non-negative. From part (*i*), we get

$$r = 1 - (\text{some non-negative quantity}) \quad \Rightarrow \quad r \le 1$$

Also from part (*ii*), we get

$$r = -1 + (\text{some non-negative quantity}) \quad \Rightarrow \quad r \ge -1 \quad \Rightarrow \quad -1 \le r$$

The sign of equality in (*) and (**) holds if and only if

$\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} = 0$ and $\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} = 0$, $(\forall\, i = 1, 2, \ldots, n)$ respectively.

$\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} = 0, \frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} = 0, \ \forall\, i = 1, 2, \ldots, n$ respectively.

From (*) and (**), we conclude that $-1 \le r \le 1$.

Example 10.4. *The variables X and Y are connected by the equation* $aX + bY + c = 0$. *Show that the correlation between them is −1 if the signs of a and b are alike and +1 if they are different.*

Solution. $aX + bY + c = 0 \quad \Rightarrow \quad aE(X) + bE(Y) + c = 0$

$$\therefore \quad a[X - E(X)] + b[Y - E(Y)] = 0 \quad \Rightarrow \quad [X - E(X)] = -\frac{b}{a}[Y - E(Y)]$$

$$\therefore \quad \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = -\frac{b}{a}E[\{Y - E(Y)\}^2] = -\frac{b}{a}\sigma_Y^2$$

and $\sigma_X^2 = E[X - E(X)]^2 = \frac{b^2}{a^2}E[\{Y - E(Y)\}^2] = \frac{b^2}{a^2} \cdot \sigma_Y^2$

$$\therefore \quad r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-\frac{b}{a} \cdot \sigma_Y^2}{\sqrt{\sigma_Y^2}\sqrt{\frac{b^2}{a^2}\sigma_Y^2}} = \frac{-\frac{b}{a}\sigma_Y^2}{\left|\frac{b}{a}\right|\sigma_Y^2}$$

$$= \begin{cases} +1, \text{ if } b \text{ and } a \text{ are of opposite signs.} \\ -1, \text{ if } b \text{ and } a \text{ are of same sign.} \end{cases}$$

Example 10.5. (*a*) *If* $Z = aX + bY$ *and r is the correlation coefficient between X and Y, show that* $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X \sigma_Y$.

(*b*) *Show that the correlation coefficient r between two random variables X and Y is given by*

$$r = (\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2) / 2\sigma_X\sigma_Y.$$

where σ_X, σ_Y and σ_{X-Y} are the standard deviations of X, Y and X − Y respectively.

Solution. (a) Taking expectation of both sides of $Z = aX + bY$, we get

$$E(Z) = aE(X) + bE(Y)$$

$$\therefore \quad Z - E(Z) = a[X - E(X)] + b[Y - E(Y)]$$

Squaring and taking expectation of both sides, we get

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \text{Cov}(X, Y) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y$$

(b) Taking $a = 1$, $b = -1$ in the above case,

$$Z = X - Y \text{ and } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$$

$$\therefore \quad r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}.$$

Remark. In the above example, we have obtained

$$V(aX + bY) = a^2\,V(X) + b^2\,V(Y) + 2ab\ \text{Cov}(X, Y)$$

Similarly, we could obtain the result $V(aX - bY) = a^2\,V(X) + b^2\,V(Y) - 2ab\ \text{Cov}(X, Y)$

The above results are useful in solving theoretical problems.

Example 10·6. *X and Y are two random variables with variances σ^2_X and σ^2_Y respectively and r is the coefficient of correlation between them. If $U = X + kY$ and $V = X + (\sigma_X/\sigma_Y)Y$, find the value of k so that U and V are uncorrelated.*

Solution. Taking expectations of $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y}Y$, we get

$$E(U) = E(X) + kE(Y) \text{ and } E(V) = E(X) + \frac{\sigma_X}{\sigma_Y}E(Y)$$

$$\therefore \quad U - E(U) = [X - E(X)] + k[Y - E(Y)]$$

and $$V - E(V) = [X - E(X)] + \frac{\sigma_X}{\sigma_Y}[Y - E(Y)]$$

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\left[\{X - E(X)\} + k\{Y - E(Y)\}\right] \times \left[\{X - E(X)\} + \frac{\sigma_X}{\sigma_Y}\{Y - E(Y)\}\right]$$

$$= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y}\text{Cov}(X, Y) + k\,\text{Cov}(X, Y) + k\frac{\sigma_X}{\sigma_Y}\cdot\sigma_Y^2$$

$$= (\sigma_X^2 + k\sigma_X\sigma_Y) + \left(\frac{\sigma_X}{\sigma_Y} + k\right)\text{Cov}(X, Y)$$

$$= \sigma_X(\sigma_X + k\sigma_Y) + \left(\frac{\sigma_X + k\sigma_Y}{\sigma_Y}\right)\text{Cov}(X, Y)$$

$$= (\sigma_X + k\sigma_Y)\left[\sigma_X + \frac{\text{Cov}(X, Y)}{\sigma_Y}\right] = (\sigma_X + k\sigma_Y)(1 + r)\sigma_X$$

U and V will be uncorrelated if $r(U, V) = 0 \quad \Rightarrow \quad \text{Cov}(U, V) = 0$

i.e., if $(\sigma_X + k\sigma_Y)(1 + r)\sigma_X = 0 \quad \Rightarrow \quad \sigma_X + k\sigma_Y = 0 \quad \Rightarrow \quad k = -\frac{\sigma_X}{\sigma_Y}$.

Example 10·7. *The random variables X and Y are jointly normally distributed and U and V are defined by $U = X\cos\alpha + Y\sin\alpha$, $V = Y\cos\alpha - X\sin\alpha$.*

Show that U and V will be uncorrelated if

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma^2_X - \sigma^2_Y},$$

where r = corr. (X, Y), σ^2_X = Var (X) and σ^2_Y = Var (Y). Are U and V then independent ?

Solution. We have : $\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$

$$= E\left[\{X - E(X)\}\cos\alpha + \{Y - E(Y)\}\sin\alpha \times [\{Y - E(Y)\}\cos\alpha - \{X - E(X)\}\sin\alpha]\right]$$

$$= \cos^2\alpha\,\text{Cov}(X, Y) - \sin\alpha\cos\alpha\,\sigma_X^2 + \sin\alpha\cos\alpha\,\sigma_Y^2 - \sin^2\alpha\,\text{Cov}(X, Y)$$

$$= (\cos^2\alpha - \sin^2\alpha)\,\text{Cov}(X, Y) - \sin\alpha\cos\alpha\,(\sigma_X^2 - \sigma_Y^2)$$

$$= \cos 2\alpha\,\text{Cov}(X, Y) - \sin\alpha\cos\alpha\,(\sigma_X^2 - \sigma_Y^2)$$

U and V will be uncorrelated if and only if $r(U, V) = 0$, *i.e.,* if $\text{Cov}(U, V) = 0$

i.e., if $\cos 2\alpha\,\text{Cov}(X, Y) - \sin\alpha\cos\alpha\,(\sigma_X^2 - \sigma_Y^2) = 0$

or if $\cos 2\alpha\, r\,\sigma_X\sigma_Y = \frac{\sin 2\alpha}{2}(\sigma_X^2 - \sigma_Y^2)$

or if $\tan 2\alpha = \frac{2r\,\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$

However, $r(U, V) = 0$, does not imply that the variables U and V are independent. [For detailed discussion, see Theorem 10·2, page 10·5.]

Example 10·8. *If X, Y are standardized random variables and*

$$r(aX + bY, bX + aY) = \frac{1 + 2ab}{a^2 + b^2}, \qquad \ldots(*)$$

find r(X, Y), the coefficient of correlation between X and Y.

Solution. Since X and Y are standardised random variables, $E(X) = E(Y) = 0$

and $\text{Var}(X) = \text{Var}(Y) = 1 \quad \Rightarrow \quad E(X^2) = E(Y^2) = 1$

and $\text{Cov}(X, Y) = E(XY) \quad \Rightarrow \quad E(XY) = r(X, Y)\,\sigma_X\sigma_Y = r(X, Y)$ $\Big\}$ $\ldots(**)$

Also we have

$$r(aX + bY, bX + aY)$$

$$= \frac{E[(aX + bY)(bX + aY)] - E(aX + bY)\,E(bX + aY)}{[\text{Var}(aX + bY)\cdot\text{Var}(bX + aY)]^{1/2}}$$

$$= \frac{E[ab\,X^2 + a^2XY + b^2YX + ab\,Y^2] - 0}{\left\{[a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)][b^2\,\text{Var}(X) + a^2\,\text{Var}(Y) + 2ba\,\text{Cov}(X, Y)]\right\}^{1/2}}$$

$$= \frac{a.b.1 + a^2r(X, Y) + b^2r(X, Y) + ab.1}{[a^2 + b^2 + 2ab\,r(X, Y)][b^2 + a^2 + 2ba\,r(X, Y)]^{1/2}} \qquad \text{[Using (**}$$

$$= \frac{2ab + (a^2 + b^2)\cdot r(X, Y)}{a^2 + b^2 + 2ab\cdot r(X, Y)}$$

From (*) and (***), we get $\quad \frac{1 + 2ab}{a^2 + b^2} = \frac{(a^2 + b^2)\cdot r(X, Y) + 2ab}{a^2 + b^2 + 2ab\cdot r(X, Y)}$

Cross multiplying, we get

$$(a^2 + b^2)(1 + 2ab) + 2ab.\,r(X, Y)(1 + 2ab) = (a^2 + b^2)^2\cdot r(X, Y) + 2ab(a^2 +$$

$$\Rightarrow \quad (a^4 + b^4 + 2a^2b^2 - 2ab - 4a^2b^2)\cdot r(X, Y) = (a^2 + b^2)$$

$$\Rightarrow \quad [(a^2 - b^2)^2 - 2ab]\,r(X, Y) = a^2 + b^2 \quad \Rightarrow \quad r(X, Y) = \frac{a^2 + b^2}{(a^2 - b^2)^2 - 2ab}.$$

Example 10-9. *If X and Y are uncorrelated random variables with means zero and variances σ_1^2 and σ_2^2 respectively, show that*

$$U = X\cos\alpha + Y\sin\alpha,\quad V = X\sin\alpha - Y\cos\alpha$$

have a correlation coefficient ρ given by $\rho = \dfrac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2\operatorname{cosec}^2 2\alpha]^{1/2}}$.

Solution. We are given that

$$r(X, Y) = 0 \Rightarrow \text{Cov}(X, Y) = 0,\quad \sigma_X^2 = \sigma_1^2 \quad \text{and} \quad \sigma_Y^2 = \sigma_2^2 \qquad \ldots(i)$$

We have

$$\begin{aligned}\sigma_U^2 &= V(X\cos\alpha + Y\sin\alpha)\\ &= \cos^2\alpha\, V(X) + \sin^2\alpha\, V(Y) + 2\sin\alpha\cos\alpha\,\text{Cov}(X, Y)\\ &= \cos^2\alpha\,\sigma_1^2 + \sin^2\alpha\,\sigma_2^2 \qquad [\text{Using } (i)]\end{aligned}$$

Similarly,

$$\sigma_V^2 = V(X\sin\alpha - Y\cos\alpha) = \sin^2\alpha\,\sigma_1^2 + \cos^2\alpha\,\sigma_2^2$$

$$\begin{aligned}\text{Cov}(U, V) &= E[\{U - E(U)\}\{V - E(V)\}]\\ &= E\big[\{(X - E(X))\cos\alpha + (Y - E(Y))\sin\alpha\} \times \{(X - E(X))\sin\alpha - (Y - E(Y))\cos\alpha\}\big]\\ &= \sin\alpha\cos\alpha\,V(X) - \cos^2\alpha\,\text{Cov}(X, Y) + \sin^2\alpha\,\text{Cov}(X, Y) - \sin\alpha\cos\alpha\,V(Y)\\ &= (\sigma_1^2 - \sigma_2^2)\sin\alpha\cos\alpha \qquad [\text{Using } (i)] \qquad \ldots(ii)\end{aligned}$$

$$\rho = \frac{\text{Cov}(U, V)}{\sigma_U\sigma_V} \quad \Rightarrow \quad \rho^2 = \frac{[\text{Cov}(U, V)]^2}{\sigma_U^2\sigma_V^2} \qquad \ldots(iii)$$

where $\sigma_U^2\sigma_V^2 = (\cos^2\alpha\,\sigma_1^2 + \sin^2\alpha\,\sigma_2^2)(\sin^2\alpha\,\sigma_1^2 + \cos^2\alpha\,\sigma_2^2)$

$$\begin{aligned}&= \sin^2\alpha\cos^2\alpha(\sigma_1^4 + \sigma_2^4) + \sigma_1^2\sigma_2^2(\cos^4\alpha + \sin^4\alpha)\\ &= \sin^2\alpha\cos^2\alpha(\sigma_1^4 + \sigma_2^4) + \sigma_1^2\sigma_2^2[(\sin^2\alpha + \cos^2\alpha)^2 - 2\sin^2\alpha\cos^2\alpha]\\ &= \sin^2\alpha\cos^2\alpha\,(\sigma_1^4 + \sigma_2^4 - 2\sigma_1^2\sigma_2^2) + \sigma_1^2\sigma_2^2\\ &= \sin^2\alpha\cos^2\alpha\,(\sigma_1^2 - \sigma_2^2)^2 + \sigma_1^2\sigma_2^2 \qquad \ldots(iv)\end{aligned}$$

$$\therefore \quad \rho^2 = \frac{(\sigma_1^2 - \sigma_2^2)^2\cdot\sin^2\alpha\cos^2\alpha}{\sigma_1^2\sigma_2^2 + \sin^2\alpha\cos^2\alpha(\sigma_1^2 - \sigma_2^2)^2} = \frac{\frac{1}{4}(\sigma_1^2 - \sigma_2^2)^2\sin^2 2\alpha}{\sigma_1^2\sigma_2^2 + \sin^2 2\alpha\cdot\frac{1}{4}(\sigma_1^2 - \sigma_2^2)^2}$$

[Using (ii) to (iv)]

$$= \frac{(\sigma_1^2 - \sigma_2^2)^2}{4\sigma_1^2\sigma_2^2\operatorname{cosec}^2 2\alpha + (\sigma_1^2 - \sigma_2^2)^2}$$

$$\Rightarrow \quad \rho = \frac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2\operatorname{cosec}^2 2\alpha]^{1/2}}.$$

Example 10-10. *If $U = aX + bY$ and $V = cX + dY$, where X and Y are measured from their respective means and if r is the correlation coefficient between X and Y, and if U and V are uncorrelated, show that* $\sigma_U\sigma_V = (ad - bc)\,\sigma_X\sigma_Y(1 - r^2)^{1/2}$.

Solution. We have

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} \quad \Rightarrow \quad 1 - r^2 = 1 - \frac{[\text{Cov}(X, Y)]^2}{\sigma_X^2\sigma_Y^2}$$

$$\Rightarrow \quad (1 - r^2)\,\sigma_X^2\sigma_Y^2 = \sigma_X^2\sigma_Y^2 - [\text{Cov}(X, Y)]^2 \qquad \ldots(*)$$

[This step is suggested by the answer.]

Since X, Y are measured from their means,

$$\left.\begin{aligned}&E(X) = 0 = E(Y) \Rightarrow E(U) = 0 = E(V)\\ &\sigma_U^2 = E(U^2), \qquad \sigma_V^2 = E(V^2)\end{aligned}\right\}$$

and

Also $\quad aX + bY - U = 0 \quad$ and $\quad cX + dY - V = 0$

$$\Rightarrow \quad \frac{X}{-bV + dU} = \frac{Y}{-cU + aV} = \frac{1}{ad - bc} \quad \text{or} \quad \left.\begin{aligned}X &= \frac{1}{ad - bc}(dU - bV)\\ Y &= \frac{1}{ad - bc}(-cU + aV)\end{aligned}\right\}$$

$$\therefore \text{Var}(X) = \frac{1}{(ad - bc)^2}[d^2\sigma_U^2 + b^2\sigma_V^2 - 2bd\,\text{Cov}(U, V)] = \frac{1}{(ad - bc)^2}[d^2\sigma_U^2 + b^2\sigma_V^2]$$

[Since U, V are uncorrelated $\Rightarrow$ Cov(U, V) = 0]

Similarly, we have

$$\text{Var}(Y) = \frac{1}{(ad - bc)^2}(c^2\sigma_U^2 + a^2\sigma_V^2)$$

$$\begin{aligned}\text{Cov}(X, Y) &= E(XY) - E(X)\,E(Y) = E(XY) \qquad [\because E(X) = 0 = E(Y)]\\ &= \frac{1}{(ad - bc)^2}E[(dU - bV)(-cU + aV)]\\ &= \frac{1}{(ad - bc)^2}[-cd\,\sigma_U^2 - ab\,\sigma_V^2] \qquad [\text{Using } (**) \text{ and Cov}(U, V) = 0]\\ &= \frac{-1}{(ad - bc)^2}[cd\,\sigma_U^2 + ab\,\sigma_V^2]\end{aligned}$$

Substituting in (*), we get

$$\begin{aligned}(1 - r^2)\,\sigma_X^2\sigma_Y^2 &= \frac{1}{(ad - bc)^4}[(d^2\sigma_U^2 + b^2\sigma_V^2)(c^2\sigma_U^2 + a^2\sigma_V^2) - (cd\sigma_U^2 + ab\sigma_V^2)^2]\\ &= \frac{1}{(ad - bc)^4}[c^2d^2\sigma_U^4 + a^2b^2\sigma_V^4 + (a^2d^2 + b^2c^2)\sigma_U^2\sigma_V^2\\ &\qquad - c^2d^2\sigma_U^4 - a^2b^2\sigma_V^4 - 2abcd\sigma_U^2\sigma_V^2]\\ &= \frac{1}{(ad - bc)^4}[a^2d^2 + b^2c^2 - 2abcd]\,\sigma_U^2\sigma_V^2 = \frac{1}{(ad - bc)^4}(ad - bc)^2\sigma_U^2\sigma_V^2\end{aligned}$$

$$= \frac{\sigma_U^2\sigma_V^2}{(ad - bc)^2}.$$

Cross multiplying and taking square root, we get the required result.

Example 10-11. *(a) Establish the formula :*

$$nr\sigma_X\sigma_Y = n_1r_1\sigma_{X_1}\sigma_{Y_1} + n_2r_2\sigma_{X_2}\sigma_{Y_2} + n_1dx_1dy_1 + n_2dx_2dy_1$$

where n_1, n_2 and n are respectively the sizes of the first, second and combined sample ($\bar{x}_1, \bar{y}_1$*), (*$\bar{x}_2, \bar{y}_2$*), (*$\bar{x}, \bar{y}$*), their means ; r_1, r_2 and r their coefficients of correl. $(\sigma_{X_1}, \sigma_{Y_1})$, $(\sigma_{X_2}, \sigma_{Y_2})$, (σ_X, σ_Y) their standard deviations, and*

$$dx_1 = \bar{x}_1 - \bar{x} \qquad dy_1 = \bar{y}_1 - \bar{y}$$

$$dx_2 = \bar{x}_2 - \bar{x} \qquad dy_2 = \bar{y}_2 - \bar{y}$$

(b) Find the correlation coefficient of combined sample given that

	Sample I	Sample II
Sample size	100	150
Sample mean $(\bar{x})$	80	72

Sample mean ($\bar{y}$)	100	118
Sample variance (σ_X^2)	10	12
Sample variance (σ_Y^2)	15	18
Correlation coefficient	0·6	0·4

(c) A set of n observations of simultaneous values of X and Y are made by an observer and standard deviations and product moment coefficient about the mean are found to be σ_X, σ_Y and ρ_{XY}. A second observer repeating the same observations made a constant error e in observing each X and a constant error E in observing each Y. The two sets of observations are combined into a single set and coefficient of correlation calculated from it. Show that its value

$$(\rho_{XY} + \tfrac{1}{4}eE) \div \sqrt{\left(\sigma_X^2 + \tfrac{1}{4}e^2\right)\left(\sigma_Y^2 + \tfrac{1}{4}E^2\right)}\ .$$

Solution. (a) Let (x_{1i}, y_{1i}); $i = 1, 2, \ldots, n_1$ and (x_{2j}, y_{2j}); $j = 1, 2, \ldots, n_2$, be the two samples of sizes n_1 and n_2 respectively from the bivariate population. Then, with the given notations, we have

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}, \qquad \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

$$\left.\begin{aligned} n\sigma_X^2 &= n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) \\ n\sigma_Y^2 &= n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) \end{aligned}\right\} \quad \ldots(i)$$

$$r_1 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1)}{n_1\sigma_{X_1}\sigma_{Y_1}}, \qquad r_2 = \frac{\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(y_{2j} - \bar{y}_2)}{n_2\sigma_{X_2}\sigma_{Y_2}} \quad \ldots(ii)$$

For the pooled sample, we have

$$r = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})(y_{2j} - \bar{y})}{n\sigma_X\sigma_Y} \quad \ldots(iii)$$

Now

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) = \sum_{i=1}^{n_1}\left[\{(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\}\{(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\}\right]$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1) + n_1(\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y})$$

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) = 0 \quad \text{and} \quad \sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1) = 0,$$

ing the algebraic sum of the deviations from the mean.

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) = n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_1\,dx_1\,dy_1 \qquad \text{[Using (ii)]}$$

Similarly, we will get

$$\sum_{j=1}^{n_2}(x_{2j} - \bar{x})(y_{2j} - \bar{y}) = n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_2\,dx_2\,dy_2$$

Substituting in (iii), we get the required formula.

(b) Here we are given :

$n_1 = 100$, $\bar{x}_1 = 80$, $\bar{y}_1 = 100$, $\sigma_{X_1}^2 = 10$, $\sigma_{Y_1}^2 = 15$, $r_1 = 0{\cdot}6$

$n_2 = 150$, $\bar{x}_2 = 72$, $\bar{y}_2 = 118$, $\sigma_{X_2}^2 = 12$, $\sigma_{Y_2}^2 = 18$, $r_2 = 0{\cdot}4$

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{100 \times 80 + 150 \times 72}{100 + 150} = 75{\cdot}2$$

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{100 \times 100 + 150 \times 118}{100 + 150} = 110{\cdot}8$$

$dx_1 = \bar{x}_1 - \bar{x} = 4{\cdot}8$, $dy_1 = \bar{y}_1 - \bar{y} = -10{\cdot}8$

$dx_2 = \bar{x}_2 - \bar{x} = 3{\cdot}2$, $dy_2 = \bar{y}_2 - \bar{y} = 7{\cdot}2$

$n\sigma_X^2 = n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) = 6640$

$n\sigma_Y^2 = n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) = 23640$

Substituting these values in the formula and simplifying, we get

$$r = \frac{n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_1 dx_1 dy_1 + n_2 dx_2 dy_2}{n\sigma_X\sigma_Y} = 0{\cdot}8186$$

(c) Here we have two sets of observations :

1st Set : (x_i, y_i), $i = 1, 2, \ldots, n$; mean $= \bar{x}$, s.d. $= \sigma_x$.

Product moment coefficient about mean $= \rho_{XY} = \text{Cov}(X, Y) = r_{xy}\sigma_x$

2nd Set : $(x_i + e, y_i + E)$, $i = 1, 2, \ldots, n$

Mean $(\bar{x}') = \frac{1}{n}\sum(x_i + e) = \bar{x} + e$

Variance $= \sigma_{x'}^2 = \frac{1}{n}\sum[(x_i + e) - (\bar{x} + e)]^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \sigma_x^2$

Mean $(\bar{y}') = \bar{y} + E$, $\sigma_{y'}^2 = \sigma_y^2$.

Product moment coefficient about mean (ρ'_{XY}) is given by :

$$\rho_{xy}' = \frac{1}{n}\sum[(x_i + e) - (\bar{x} + e)][(y_i + E) - (\bar{y} + E)] = \rho_{xy}$$

To obtain the correlation coefficient for the combined set of 2n observa Formula (10·5) in Part (a) above.

Example 10·12. *The independent variables X and Y are defined by :*

$$f(x) = \begin{cases} 4ax, & 0 \le x \le r \\ 0, & \text{otherwise} \end{cases} \qquad \phi(y) = \begin{cases} 4by, & 0 \le y \le s \\ 0, & \text{otherwise} \end{cases}$$

Show that : Corr. $(U, V) = \frac{b - a}{b + a}$, *where* $U = X + Y$ *and* $V = X - Y$

Solution. Since the total area under probability curve is unity, we have

$$\int_0^r f(x)\,dx = 4a\int_0^r x\,dx = 1 \quad \Rightarrow \quad 2a\,r^2 = 1 \quad \Rightarrow \quad a = \frac{1}{2r^2} \qquad \ldots(1)$$

$$\int_0^s f(y)\,dy = 4b\int_0^s y\,dy = 1 \quad \Rightarrow \quad 2bs^2 = 1 \quad \Rightarrow \quad b = \frac{1}{2s^2} \qquad \ldots(2)$$

$$\therefore \quad f(x) = 4ax = \frac{2x}{r^2},\ 0 \le x \le r; \text{ and } f(y) = 4by = \frac{2y}{s^2},\ 0 \le y \le s \qquad \ldots(3)$$

Since X and Y are independent variates,

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}(X, Y) = 0 \qquad \ldots(4)$$

$$\text{Cov}(U, V) = \text{Cov}(X + Y, X - Y)$$
$$= \text{Cov}(X, X) - \text{Cov}(X, Y) + \text{Cov}(Y, X) - \text{Cov}(Y, Y)$$
$$= \sigma_X^2 - \sigma_Y^2 \qquad \text{[Using (4)]}$$

$$\text{Var}(U) = \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$$
$$= \sigma_X^2 + \sigma_Y^2 \qquad \text{[Using (4)]}$$

$$\text{Var}(V) = \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y)$$
$$= \sigma_X^2 + \sigma_Y^2 \qquad \text{[Using (4)]}$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \qquad \ldots(5)$$

$$\therefore \quad E(X) = \int_0^r x f(x)\,dx = \frac{2}{r^2}\int_0^r x^2\,dx = \frac{2r}{3} \qquad \text{[From (3)]}$$

$$E(X^2) = \int_0^r x^2 f(x)\,dx = \frac{2}{r^2}\int_0^r x^3\,dx = \frac{r^2}{2}$$

$$\Rightarrow \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{r^2}{2} - \frac{4r^2}{9} = \frac{r^2}{18} = \frac{1}{36a} \qquad \text{[From (1)]}$$

Similarly, we shall get

$$E(Y) = \frac{2s}{3},\ E(Y^2) = \frac{s^2}{2} \quad \text{and} \quad \text{Var}(Y) = \frac{s^2}{18} = \frac{1}{36b}$$

Substituting in (5), we get

$$r(U, V) = \frac{(1/36a) - (1/36b)}{(1/36a) + (1/36b)} = \frac{b - a}{b + a}$$

Example 10-13. *Let the random variable X have the marginal density :*

$$f_1(x) = 1, -\tfrac{1}{2} < x < \tfrac{1}{2},$$

and let the conditional density of Y be

$$f(y \mid x) = \begin{cases} 1, x < y < x + 1, & -\frac{1}{2} < x < 0 \\ 1, -x < y < 1 - x, & 0 < x < \frac{1}{2} \end{cases} \qquad \ldots(*)$$

Show that the variables X and Y are uncorrelated.

Solution. We have

$$E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x f_1(x)\,dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x \cdot 1\,dx = \left|\frac{x^2}{2}\right|_{-\frac{1}{2}}^{\frac{1}{2}} = 0$$

If $f(x, y)$ is the joint p.d.f. of X and Y, then

$$f(x, y) = f(y \mid x) f_1(x) = f(y \mid x) \qquad [\because f_1(x) = 1]$$

$$E(XY) = \int_{-\frac{1}{2}}^{0}\int_{x}^{x+1} xy\,dx\,dy + \int_{0}^{\frac{1}{2}}\int_{-x}^{1-x} xy\,dx\,dy$$

$$= \int_{-\frac{1}{2}}^{0}\left(x\int_{x}^{x+1} y\,dy\right)dx + \int_{0}^{\frac{1}{2}}\left(x\int_{-x}^{1-x} y\,dy\right)dx \qquad \text{[From (*)]}$$

$$= \frac{1}{2}\int_{-\frac{1}{2}}^{0} x(2x + 1)\,dx + \frac{1}{2}\int_{0}^{\frac{1}{2}} x(1 - 2x)\,dx$$

$$= \frac{1}{2}\left[\frac{2}{3}x^3 + \frac{x^2}{2}\right]_{-1/2}^{0} + \frac{1}{2}\left[\frac{1}{2}x^2 - \frac{2}{3}x^3\right]_{0}^{1/2}$$

$$= \frac{1}{2}\left(\frac{1}{12} - \frac{1}{8} - \frac{1}{12} + \frac{1}{8}\right) = 0$$

$$\therefore \quad \text{Cov}(XY) = E(XY) - E(X)\,E(Y) = 0 \quad \Rightarrow \quad r(X, Y) = 0$$

Hence the variables X and Y are uncorrelated.

10·5. CALCULATION OF THE CORRELATION COEFFICIENT FOR A BI… FREQUENCY DISTRIBUTION

When the data are considerably large, they may be summarised by usi… way table. Here, for each variable a suitable number of classes are taken, k… view the same considerations as in the univariate case. If there are m classes … n classes for Y, there will be in all $m \times n$ cells in the two-way table. By going … the pairs of values of X and Y, we can find the frequency for each cell. The … of cell frequencies will then define a *bivariate frequency distribution*. The colu… and row totals will give us the marginal distributions of X and Y respe… particular column or row will be called the conditional distribution of Y for … of X for given Y respectively.

Suppose that the bivariate data on X and Y are presented in a two-way … table (shown on *page 10·18*) where there are n classes of Y placed along the … lines and m classes of X along the vertical lines and f_{ij} is the frequency of … lying in the (i, j)th cell.

Here $\sum_x f(x, y) = g(y)$, is the sum of the frequencies along any row an…

$\sum_y f(x, y) = f(x)$ is the sum of the frequencies along any column.

We observe that $\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$

$$\therefore \quad \bar{x} = \frac{1}{N}\sum_x\sum_y x f(x,y) = \frac{1}{N}\left[\sum_x \left\{x \sum_y f(x,y)\right\}\right] = \frac{1}{N}\sum_x x f(x)$$

Similarly

$$\bar{y} = \frac{1}{N}\sum_x\sum_y y f(x,y) = \frac{1}{N}\sum_y y\, g(y)$$

$$\sigma_X^2 = \frac{1}{N}\sum_x\sum_y x^2 f(x,y) - \bar{x}^2 = \frac{1}{N}\sum_x x^2 f(x) - \bar{x}^2$$

$$\sigma_Y^2 = \frac{1}{N}\sum_x\sum_y y^2 f(x,y) - \bar{y}^2 = \frac{1}{N}\sum_y y^2 g(y) - \bar{y}^2$$

BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)

Y Series ↓ \ X Series →	Classes: Mid Points x_1 x_2 ... x_i ... x_m	Total of Frequencies of Y $g(y)$
y_1		
y_2		
⋮		
y_j	$f(x, y)$	$g(y) = \sum_x f(x, y)$
⋮		
y_n		
Total of frequencies of X $f(x)$	$f(x) = \sum_y f(x, y)$	$N \to \sum_x\sum_y f(x, y)$ ↓ $\sum_y\sum_x f(x, y)$

Example 10·14. *The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test :*

Ages in years → / Marks ↓	18	19	20	21	Total
10—20	4	2	2	—	8
20—30	5	4	6	4	19
30—40	6	8	10	11	35
40—50	4	4	6	8	22
50—60	—	2	4	4	10
60—70	—	2	3	1	6
Total	19	22	31	28	100

Calculate the correlation coefficient.

Solution.

CORRELATION TABLE

v	Mid-value	u : Age (X) / Marks (Y)	−1 : 18	0 : 19	1 : 20	−2 : 21	Total $g(v)$	$vg(v)$	$v^2g(v)$	$\sum_u uvf(u,v)$
−2	15	10—20	(8) 4	(0) 2	(−4) 2		8	−16	32	4
−1	25	20—30	(5) 5	(0) 4	(−6) 6	(−8) 4	19	−19	19	−9
0	35	30—40	(0) 6	(0) 8	(0) 10	(0) 11	35	0	0	0
1	45	40—50	(−4) 4	(0) 4	(6) 6	(16) 8	22	22	22	18
2	55	50—60		(0) 2	(8) 4	(16) 4	10	20	40	24
3	65	60—70		(0) 2	(9) 3	(6) 1	6	18	54	15
		Total $f(u)$	19	22	31	28	100	25	167	52
		$uf(u)$	−19	0	31	56	68			
		$u^2f(u)$	19	0	31	112	162			
		$\sum_v uvf(u,v)$	9	0	13	30	52			

Let $U = X - 19, \quad V = |(Y - 35)/10|$

$$\bar{u} = \frac{1}{N}\sum_u uf(u) = \frac{68}{100} = 0.68, \quad \bar{v} = \frac{1}{N}\sum_v v\, g(v) = \frac{25}{100} = 0.25$$

$$\text{Cov}(u, v) = \frac{1}{N}\sum_u\sum_v uv\, f(u,v) - \bar{u}\,\bar{v} = \frac{1}{100}\times 52 - 0.68 \times 0.25 = 0.35$$

$$\sigma_U^2 = \frac{1}{N}\sum_u u^2 f(u) - \bar{u}^2 = \frac{162}{100} - (0.68)^2 = 1.1576$$

$$\sigma_V^2 = \frac{1}{N}\sum_v v^2 g(v) - \bar{v}^2 = \frac{167}{100} - (0.25)^2 = 1.6075$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U\,\sigma_V} = \frac{0.35}{\sqrt{1.1576 \times 1.6075}} = 0.25$$

Since correlation coefficient is independent of change of origin and scale

$$r(X, Y) = r(U, V) = 0.25.$$

Remark. Figures in circles in the table are the product terms $uvf(u, v)$.

Example 10·15. *The joint probability distribution of X and Y is given below :*

Y \ X	−1	+1
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$

Find the correlation coefficient between X and Y.

Solution.

COMPUTATION OF MARGINAL PROBABILITIES

Y \ X	−1	+1	g(y)
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
p(x)	$\frac{3}{8}$	$\frac{5}{8}$	1

We have

$$E(X) = \Sigma x p(x) = (-1)\times\frac{3}{8} + 1\times\frac{5}{8} = \frac{1}{4}, \qquad E(X^2) = \Sigma x^2 p(x) = (-1)^2\times\frac{3}{8} + 1^2\times\frac{5}{8} = 1$$

$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = 1 - \frac{1}{16} = \frac{15}{16}$$

$$E(Y) = \Sigma y\, g(y) = 0\times\frac{4}{8} + 1\times\frac{4}{8} = \frac{1}{2}, \qquad E(Y^2) = \Sigma y^2 g(y) = 0^2\times\frac{4}{8} + 1^2\times\frac{4}{8} = \frac{1}{2}$$

$$\therefore \quad \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$E(XY) = 0\times(-1)\times\frac{1}{8} + 0\times 1\times\frac{3}{8} + 1\times(-1)\times\frac{2}{8} + 1\times 1\times\frac{2}{8} = -\frac{2}{8} + \frac{2}{8} = 0$$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = 0 - \frac{1}{4}\times\frac{1}{2} = -\frac{1}{8}$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-\frac{1}{8}}{\sqrt{\frac{15}{16}\times\frac{1}{4}}} = \frac{-1}{\sqrt{15}} = \frac{-1}{3.873} = -0.2582.$$

Example 10·16. *If X_1, X_2 and X_3 are uncorrelated variables with equal mean M and variances V_1^2, V_2^2 and V_3^2 respectively, prove that correlation coefficient ρ between $Z_1 = \frac{X_1}{X_3}$ and $Z_2 = \frac{X_2}{X_3}$ is given by ;* $\rho = \dfrac{V_3^2}{\sqrt{(V_1^2 + V_3^2)(V_2^2 + V_3^2)}}$.

Solution. Let us make N observations on each of the variables X_1, X_2 and X_3. Then are given :

$$\overline{X}_1 = \overline{X}_2 = \overline{X}_3 = M$$

Let x_i be the deviation of X_i, $(i = 1, 2, 3)$ from the mean M so that

$$x_{ij} = X_{ij} - M \quad \Rightarrow \quad X_{ij} = x_{ij} + M \; (i = 1, 2, 3); j = 1, 2, \ldots, N \qquad \ldots (i)$$

$$\therefore \quad \sum_{j=1}^{N} x_{ij} = \sum_{j=1}^{N} (X_{ij} - M) = 0; \; i = 1, 2, 3$$

being the algebraic sum of the deviations of the observations from their mean.

Also $\quad V_i^2 = \text{Var}(X_i) = \frac{1}{N}\sum_j (X_{ij} - M)^2 = \frac{1}{N}\sum_j x_{ij}^2 \; ; i = 1, 2, 3$

Further, since X_1, X_2 and X_3 are uncorrelated, we have

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, X_3) = \text{Cov}(X_2, X_3) = 0$$

$$\Rightarrow \quad \sum_j x_{1j} x_{2j} = \sum_j x_{1j} x_{3j} = \sum_j x_{2j} x_{3j} = 0$$

$$Z_1 = \frac{X_1}{X_3} \quad \text{and} \quad Z_2 = \frac{X_2}{X_3} \qquad \text{(Given)}$$

Let m_1, σ_1 be the mean and *s.d.* respectively of Z_1 and (m_2, σ_2) be the mean respectively of Z_2. Then, we get

$$m_1 = \overline{Z}_1 = \frac{1}{N}\sum_{j=1}^{N}\frac{X_{1j}}{X_{3j}} = \frac{1}{N}\sum_j\left(\frac{x_{1j} + M}{x_{3j} + M}\right)$$

$$= \frac{1}{N}\sum_j\left[\left(1 + \frac{x_{1j}}{M}\right)\left(1 + \frac{x_{3j}}{M}\right)^{-1}\right]$$

$$= \frac{1}{N}\sum_j\left[\left(1 + \frac{x_{1j}}{M}\right)\left(1 - \frac{x_{3j}}{M} + \frac{x_{3j}^2}{M^2} - \frac{x_{3j}^3}{M^3} + \ldots\right)\right]$$

$$= \frac{1}{N}\sum_j\left(1 + \frac{x_{1j}}{M} - \frac{x_{3j}}{M} - \frac{x_{1j}x_{3j}}{M^2} + \frac{x_{3j}^2}{M^2}\right),$$

neglecting $(x/M)^3$ and higher powers of (x/M).

$$\therefore \quad m_1 = 1 + \frac{V_3^2}{M^2} \qquad \text{[From (ii)}$$

Similarly, we shall get $\quad m_2 = \overline{Z}_2 = 1 + \frac{V_3^2}{M^2}$

$$\therefore \quad m_1 = m_2 = 1 + \frac{V_3^2}{M^2}$$

$$\sigma_1^2 = \text{Var}(Z_1) = \frac{1}{N}\sum_j\left(\frac{X_{1j}}{X_{3j}}\right)^2 - m_1^2$$

$$\frac{1}{N}\sum_j\left(\frac{X_{1j}}{X_{3j}}\right)^2 = \frac{1}{N}\sum_j (x_{1j} + M)^2 (x_{3j} + M)^{-2} = \frac{1}{N}\sum_j\left(1 + \frac{x_{1j}}{M}\right)^2\left(1 + \frac{x_{3j}}{M}\right)^{-2}$$

$$= \frac{1}{N}\sum_j\left[\left(1 + \frac{x_{1j}^2}{M^2} + \frac{2x_{1j}}{M}\right)\left\{1 - \frac{2x_{3j}}{M} + \frac{(-2)(-3)}{2!}\left(\frac{x_{3j}}{M}\right)^2\right\}\right]$$

(neglecting higher powers of (x/M).)

$$= \frac{1}{N}\sum_j\left[1 + \frac{x_{1j}^2}{M^2} + \frac{2x_{1j}}{M} - \frac{2x_{3j}}{M} - 4\frac{x_{1j}x_{3j}}{M^2} + 3\frac{x_{3j}^2}{M^2}\right],$$

neglecting higher powers of (x/M).

$$= 1 + \frac{V_1^2}{M^2} + \frac{3V_3^2}{M^2} \quad \text{[From (ii), (iii) and (iv)]} \qquad \ldots (vii)$$

Substituting from (vii) in (vi), we get

$$s_1^2 = 1 + \frac{V_1^2}{M^2} + \frac{3V_3^2}{M^2} - \left(1 + \frac{V_3^2}{M^2}\right)^2$$

$$= 1 + \frac{V_1^2}{M^2} + \frac{3V_3^2}{M^2} - \left(1 + \frac{2V_3^2}{M^2} + \frac{V_3^4}{M^4}\right)$$

$$= \frac{V_1^2}{M^2} + \frac{V_3^2}{M^2} = \frac{1}{M^2}(V_1^2 + V_3^2), \qquad \ldots (viii)$$

her terms being neglected.

Similarly, we shall get : $$s_2^2 = \frac{1}{M^2}(V_2^2 + V_3^2) \qquad \ldots (ix)$$

$$\text{Cov}(Z_1, Z_2) = \frac{1}{N}\sum_i \left(\frac{X_{1i}}{X_{3i}} \cdot \frac{X_{2i}}{X_{3i}}\right) - m_1 m_2$$

$$= \frac{1}{N}\sum_i (x_{1i} + M)(x_{2i} + M)(x_{3i} + M)^{-2} - m_1 m_2$$

$$= \frac{1}{N}\sum_i \left(1 + \frac{x_{1i}}{M}\right)\left(1 + \frac{x_{2i}}{M}\right)\left(1 + \frac{x_{3i}}{M}\right)^{-2} - m_1 m_2$$

$$= \frac{V_3^2}{M^2} \quad \text{[On simplifying and using (ii) to (v)]} \qquad \ldots (x)$$

The correlation coefficient ρ between Z_1 and Z_2 is given by :

$$\rho = \frac{\text{Cov}(Z_1 Z_2)}{\sigma_{Z_1}\sigma_{Z_2}} = \frac{V_3^2}{M^2} \cdot \frac{M^2}{\sqrt{(V_1^2 + V_3^2)(V_2^2 + V_3^2)}} \quad \text{[From (viii), (ix) and (x)]}$$

$$= \frac{V_3^2}{\sqrt{(V_1^2 + V_3^2)(V_2^2 + V_3^2)}}.$$

10.6. PROBABLE ERROR OF CORRELATION COEFFICIENT

If r is the correlation coefficient in a sample of n pairs of observations, then its standard error (S.E.) is given by : $\text{S.E.}(r) = \frac{1 - r^2}{\sqrt{n}}$.

Probable error (P.E.) of correlation coefficient is given by :

$$\text{P.E.}(r) = 0.6745 \times \text{S.E.}(r) = 0.6745 \frac{(1 - r^2)}{\sqrt{n}} \qquad \ldots (10\text{-}6)$$

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0·6745 is that in a normal distribution, the range $\mu \pm 0.6745\,\sigma$ covers 50% of the total area. According to Secrist, *"The probable error of the correlation coefficient is an amount which if added to and subtracted from the mean correlation coefficient, produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall."*

all significant. If $r > 6$ P.E.(r), it is definitely

significant. A rigorous method (t-test) of testing the significance of an observed correlation coefficient will be discussed later in *"tests of significance"* in sampling.

Probable error also enables us to find the limits within which the population correlation coefficient can be expected to vary. The limits are $r \pm \text{P.E.}(r)$.

10.7. RANK CORRELATION

Let us suppose that a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let (x_i, y_i); $i = 1, 2, \ldots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the *rank correlation coefficient* between A and B for that group of individuals.

10·7·1. Spearman's Rank Correlation Coefficient. Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values 1, 2, …, n.

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

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}(1^2 + 2^2 + \ldots + n^2) - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2 - 1}{12}$$

$$\therefore \quad \sigma_X^2 = \frac{n^2 - 1}{12} = \sigma_Y^2 \qquad \ldots (*)$$

In general $x_i \neq y_i$. Let $d_i = x_i - y_i$ $\quad \therefore \quad d_i = (x_i - \bar{x}) - (y_i - \bar{y})$

Squaring and summing over i from 1 to n, we get

$$\sum_i d_i^2 = \sum_i [(x_i - \bar{x}) - (y_i - \bar{y})]^2 = \Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\Sigma d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\,\sigma_X\sigma_Y$$

where ρ is the rank correlation coefficient between A and B.

$$\therefore \quad \frac{1}{n}\Sigma d_i^2 = 2\sigma_X^2 - 2\rho\sigma_X^2 \quad \Rightarrow \quad 1 - \rho = \frac{\Sigma d_i^2}{2n\sigma_X^2}$$

$$\Rightarrow \quad \rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \qquad \text{[From (*)]}$$

which is the *Spearman's formula for the rank correlation coefficient.*

Remark. We always have $\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = n(\bar{x} - \bar{y}) = 0$

This serves as a check on the calculations.

10·7·2. Tied Ranks. If some of the individuals receive the same rank in merit, they are said to be tied. Let us suppose that m of the individuals, s $(k+2)^{th}$, ..., $(k+m)^{th}$ are tied. Then each of these m individuals is assigne rank, which is arithmetic mean of the ranks $k+1, k+2, \ldots, k+m$.

Derivation of $\rho(X, Y)$. We have :

$$\rho(X, Y) = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{[\Sigma (X - \bar{X})^2 . \Sigma (Y - \bar{Y})^2]^{1/2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \ \Sigma y^2}}, \qquad ...(i)$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$.

If X and Y each takes the values $1, 2, ..., n$, then $\bar{X} = (n + 1)/2 = \bar{Y}$

and $$n\sigma_X^2 = \Sigma x^2 = \frac{n(n^2 - 1)}{12} \text{ and } n\sigma_Y^2 = \Sigma y^2 = \frac{n(n^2 - 1)}{12} \qquad ...(ii)$$

Also $$\Sigma d^2 = \Sigma (X - Y)^2 = \Sigma [(X - \bar{X}) - (Y - \bar{Y})]^2 = \Sigma (x - y)^2$$

$$\Rightarrow \quad \Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy \quad \Rightarrow \quad \Sigma xy = \frac{1}{2}\left(\Sigma x^2 + \Sigma y^2 - \Sigma d^2\right) \qquad ...(iii)$$

We shall now investigate the effect of common ranking (in case of ties), on the sum of squares of the ranks. Let S^2 and S_1^2 denote the sum of the squares of untied and tied ranks respectively. Then we have :

$$S^2 = (k + 1)^2 + (k + 2)^2 + ... + (k + m)^2$$
$$= mk^2 + (1^2 + 2^2 + ... + n^2) + 2k (1 + 2 + ... + n)$$
$$= mk^2 + \frac{n (n + 1) (2n + 1)}{6} + nk (n + 1)$$

$$S_1^2 = m \text{ (Average rank)}^2 = m \left[\frac{(k + 1) + (k + 2) + ... + (k + m)}{m} \right]^2$$

$$= m\left(k + \frac{m + 1}{2}\right)^2 = mk^2 + \frac{m (m + 1)^2}{4} + mk (m + 1)$$

$$\therefore \quad S^2 - S_1^2 = \frac{m(m + 1)}{12}\left[2 (2m + 1) - 3 (m + 1)\right] = \frac{m (m^2 - 1)}{12}$$

Thus the effect of tying m individuals (ranks) is to reduce the sum of the squares by $m (m^2 - 1)/12$, though the mean value of the ranks remains the same, *viz.*, $(n + 1)/2$.

Suppose that there are s such sets of ranks to be tied in the X-series so that the total sum of squares due to them is :

$$\frac{1}{12} \sum_{i=1}^{s} m_i (m_i^2 - 1) = \frac{1}{12} \sum_{i=1}^{s} (m_i^3 - m_i) = T_X, \text{ (say)} \qquad ... (10·7a)$$

Similarly suppose that there are t such sets of ranks to be tied with respect to the other series Y so that sum of squares due to them is :

$$\frac{1}{12} \sum_{j=1}^{t} m_j' (m_j'^2 - 1) = \frac{1}{12} \sum_{j=1}^{t} (m_j'^3 - m_j') = T_Y, \text{ (say)} \qquad ...(10·7b)$$

Thus, in the case of ties, the new sums of squares are given by :

$$n \text{ Var}'(X) = \Sigma x^2 - T_X = \frac{n(n^2 - 1)}{12} - T_X, \qquad n \text{ Var}'(Y) = \Sigma y^2 - T_Y = \frac{n(n^2 - 1)}{12} - T_Y$$

and $$n \text{ Cov}'(X, Y) = \frac{1}{2}(\Sigma x^2 - T_X + \Sigma y^2 - T_Y - \Sigma d^2)$$

$$= \frac{1}{2}\left[\frac{n(n^2 - 1)}{12} - T_X + \frac{n (n^2 - 1)}{12} - T_Y - \Sigma d^2\right]$$

$$= \frac{n(n^2 - 1)}{12} - \frac{1}{2}\left[(T_X + T_Y) + \Sigma d^2\right]$$

$$\rho(X, Y) = \frac{\frac{n (n^2 - 1)}{12} - \frac{1}{2}(T_X + T_Y + \Sigma d^2)}{\left[\frac{n(n^2 - 1)}{12} - T_X\right]^{1/2}\left[\frac{n (n^2 - 1)}{12} - T_Y\right]^{1/2}}$$

$$= \frac{[n(n^2 - 1)/6] - (\Sigma d^2 + T_X + T_Y)}{\left[\frac{n(n^2 - 1)}{6} - 2T_X\right]^{1/2}\left[\frac{n(n^2 - 1)}{6} - 2T_Y\right]^{1/2}} \qquad ...(10·7c)$$

where T_X and T_Y are given by (10·7*a*) and (10·7*b*).

Remark. If we adjust only the covariance term, *i.e.*, Σxy and not the variances σ_x^2 (or Σx^2) and σ_y^2 (or Σy^2) for ties, then the formula (10·7*c*) reduces to :

$$\rho(X, Y) = \frac{[n (n^2 - 1)/6] - (\Sigma d^2 + T_X + T_Y)}{n(n^2 - 1)/6} = 1 - \frac{6 (\Sigma d^2 + T_X + T_Y)}{n(n^2 - 1)} \qquad ...(10·7d)$$

a formula which is commonly used in practice for numerical problems. For illustration, see Example 10·19.

Example 10·17. *The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics :*

(1, 1) (2, 10) (3, 3) (4, 4) (5, 5) (6, 7) (7, 2) (8, 6) (9, 8) (10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13).

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

Solution.

Ranks in Maths. (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	−8	0	0	0	−1	5	2	1	−1	−4	3	−1	2	−1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by :

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0·8.$$

Example 10·18. *Ten competitors in a musical test were ranked by the three judges A, B and C in the following order :*

Ranks by A	:	1	6	5	10	3	2	4	9	7
Ranks by B	:	3	5	8	4	7	10	2	1	6
Ranks by C	:	6	4	9	8	1	2	3	10	5

(From (10))

...(10·7c)

riances σ_x^2 (or $\sum x^2$)

...(10·7d)

r illustration, see

d Physics are as
Mathematics and

0, 11) (11, 15)

n Mathematics

15	16	Total
6	13	
1	3	0
	9	136

judges A, B

7	8
6	9
5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Solution. Here $n = 10$

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−5	−3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	−3	−4	−1	9	16	1
10	4	8	6	2	−4	36	4	16
3	7	1	−4	2	6	16	4	36
2	10	2	−8	0	8	64	0	64
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	−1	1	2	1	1	4
Total			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 60$	$\Sigma d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = 1 - \frac{214}{165} = -\frac{49}{165}.$$

Since $\rho(X, Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

10·7·3. Repeated Ranks (Continued). If any two or more individuals are ticketed equal in any classification with respect to characteristics A and B, or if there more than one item with the same value in the series, then the Spearman's formula 7) for calculating the rank correlation coefficient breaks down, since in this case h of the variables X and Y does not assume the values 1, 2, ..., n and consequently, $\bar{Y}$.

In this case, common ranks are given to the repeated items. This common rank is average of the ranks which these items would have assumed if they were slightly ferent from each other and the next item will get the rank next to the ranks already igned. As a result of this, following adjustment or correction is made in the rank relation formula [see (10·7d)].

In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m is the number of es an item is repeated. This correction factor is to be added for each repeated value oth the X-series and Y-series.

Example 10·19. *Obtain the rank correlation coefficient for the following data :*

X :	68	64	75	50	64	80	75	40	55	64
Y :	62	58	68	45	81	60	68	48	50	70

Solution.

CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	$d = x - y$	d^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
				$\Sigma d = 0$	$\Sigma d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To $\sum d^2$ we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is to be applied twice, once for the value 75 which occurs twice ($m = 2$) and then for the value 64 which occurs thrice ($m = 3$). The total correction for the X-series is : $\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$. Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$, as the value 68 occurs twice.

$$\therefore \quad \rho = 1 - \frac{6\left(\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right)}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0.545.$$

10·7·4. Limits for the Rank Correlation Coefficient. Spearman's Rank correlation coefficient is given by : $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$.

'ρ' is maximum, if $\sum_{i=1}^{n} d_i^2$ is minimum, *i.e.*, if each of the deviations d_i is minimum But the minimum value of d_i is zero in the particular case $x_i = y_i$, *i.e.*, if the ranks of the ith individual in the two characteristics are equal. Hence the maximum value of

+1, i.e., $\rho \le 1$. 'ρ' is minimum, if $\sum_{i=1}^{n} d_i^2$ is maximum, i.e., if each of the deviations

maximum which is so if the ranks of the n individuals in the two characteristics are in the opposite directions as given below :

x	1	2	3	...	...	$n-1$	n
y	n	$n-1$	$n-2$	...	...	2	1

... (*)

Case I. Suppose n is odd equal to $(2m + 1)$, then the values of d are :

$d : 2m, 2m-2, 2m-4, \ldots, 2, 0, -2, -4, \ldots -(2m-2), -2m.$

$$\therefore \sum_{i=1}^{n} d_i^2 = 2[(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2]$$

$$= 8\left[m^2 + (m-1)^2 + \ldots + 2^2 + 1^2\right] = \frac{8m(m+1)(2m+1)}{6}$$

Hence $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2-1\}} = 1 - \frac{8m(m+1)}{4m(m+1)} = -1.$

Case II. Let n be even and equal to $2m$, (say).

Then the values of d are : $(2m-1), (2m-3), \ldots, 1, -1, -3, \ldots, -(2m-3), -(2m-1)$

$$\therefore \quad \Sigma d_i^2 = 2\{(2m-1)^2 + (2m-3)^2 + \ldots + 1^2\}$$

$$= 2[\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + \ldots + 2^2 + 1^2\}$$

$$-\{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}]$$

$$= 2[1^2 + 2^2 + \ldots + (2m)^2 - \{2^2m^2 + 2^2(m-1)^2 + \ldots + 2^2\}]$$

$$= 2\left[\frac{2m(2m+1)(4m+1)}{6} - 4\frac{m(m+1)(2m+1)}{6}\right]$$

$$= \frac{2m}{3}[(2m+1)(4m+1) - 2(m+1)(2m+1)]$$

$$= \frac{2m}{3}[(2m+1)(4m+1-2m-2)] = \frac{2m}{3}(2m+1)(2m-1) = \frac{2m(4m^2-1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{4m(4m^2-1)}{2m(4m^2-1)} = -1$$

Thus the limits for rank correlation coefficient are given by $-1 \le \rho \le 1$.

Aliter. For an alternate and simpler proof for obtaining the minimum value of ρ, from (*) onward, proceed as in Hint to *Question Number 10.51* of Assorted Review Problems for Self-Assessment.

Remarks on Spearman's Rank Correlation Coefficient.

1. $\Sigma d = \Sigma x - \Sigma y = n(\bar{x} - \bar{y}) = 0$, which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated, then we need a measure which is distribution free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the population from which sample observations are drawn.

4. Spearman's formula is easy to understand and apply as compared with Karl Pearson's formula. The value obtained by the two formulae, viz., Pearsonian r and Spearman's ρ are generally different. The difference arises due to the fact that when ranking is used instead of set of observations, there is always some loss of information. Unless many ties are there, coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient when we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of large observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

CHAPTER CONCEPTS QUIZ

1. State whether the following statements are TRUE or FALSE :

(i) $r_{XY} = 0 \Rightarrow X$ and Y are independent.

(ii) If $r_{XY} > 0$, then $r_{X,-Y} > 0$, $r_{-X,Y} > 0$ and $r_{-X,-Y} > 0$.

(iii) $r_{XY} > 0 \Rightarrow E(XY) > E(X)\,E(Y)$.

(iv) Pearson's coefficient of correlation is independent of origin but not of scale.

(v) The numerical value of product moment correlation coefficient 'r' between variables X and Y cannot exceed unity.

(vi) If the correlation coefficient between the variables X and Y is zero the correlation coefficient between X^2 and Y^2 is also zero.

(vii) If $r > 0$, then as X increases, Y also increases.

(viii) r measures every type of relationship between the two variables.

(ix) "High positive coefficient of correlation between increase in the sale of newspapers and increase in the number of crimes leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes.

(x) "A high positive value of r between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer.

(xi) Let (X, Y) be jointly discrete random variables such that each X and Y has two mass points. Then X and Y are independent if and only if they are uncorrelated.

2. Fill in the blanks :

(i) If $r_{XY} = 1$, the line on the graph will extend from ... to ...

(ii) If $r_{XY} = 0$, it depicts ... association.

(iii) Correlation coefficient is a ... number.

(iv) If r is the correlation coefficient, the $\sqrt{1-r^2}$ is termed as ... .

(v) The Karl Pearson coefficient of correlation between variables X and Y is ...

(vi) Two independent variables are .....

(vii) Limits for correlation coefficient are .....

(viii) If r be the correlation coefficient between the random variables X and Y the variance of $X + Y$ is .....

10-53. If d_i be the difference in the ranks of the ith individual in two different characteristics, then show that the maximum value of $\sum_{i=1}^{n} d_i^2$ is $\frac{1}{3}(n^3 - n)$. Hence or otherwise, show that rank correlation coefficient lies between -1 and $+1$.

10-54. Let $x_1, x_2, \ldots, x_n$ be the ranks of n individuals according to a character A and $y_1, y_2, \ldots,$ be the corresponding ranks of the same individuals according to other character B. Obviously $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are permutations of $1, 2, \ldots, n$. It is given that $x_i + y_i = n$, for $i = 1, 2, \ldots, n$. Show that the value of the rank correlation coefficient ρ between the characters A and B is -1.

10-55. Show that in the ranked bivariate distribution in which no ties occur and in which the variables are independent.

(*i*) $\sum_i d_i^2$ is always even, and (*ii*) there are not more than $\left[\frac{1}{6}(n^3 - n) + 1\right]$ values of r.

10-56. Each of n independent trials can materialise in exactly one of the results $A_1, A_2, \ldots, A_k$. the probability of A_i is p_i in every trial, $\left(\sum_{i=1} p_i = 1\right)$, find the probability of obtaining the equencies $r_1, r_2, \ldots, r_k$ for $A_1, A_2, \ldots, A_k$ respectively in these trials. Also find $E(r_i)$, Var (r_i) and ow that the correlation coefficient between r_i and r_j is independent of n.

10-57. In a sample of size n from a multinomial population $n_1, n_2, \ldots, n_k$ are of type $1, 2, \ldots, k$ th $\sum p_i = 1$, where p_i is the probability of type i $(i = 1, 2, \ldots, k)$. Show that the expected value of when n_1 is given is $(n - n_1)\, p_2 (1 - p_1)$ and hence or otherwise show that the coefficient of rrelation between n_i and n_j is :

$$-\left[\frac{p_i p_j}{(1-p_i)(1-p_j)}\right]^{\frac{1}{2}}$$

10-58. A ball is drawn at random from an urn containing 3 white balls numbered 0, 1, 2 ; 2 balls numbered 0, 1 and 1 back ball numbered 0. If the colours white, red and black are ain numbered 0, 1 and 2 respectively, show that the correlation coefficient between the riables : X, the colour number and Y, the number of the ball is $-\frac{1}{2}$.

10-59. A fair coin is tossed three times. Let X = number of heads in three tossings and difference, in absolute value, between number of heads and number of tails. Determine the t *p.m.f.* of (X, Y) and also the two marginal *p.m.f's*. Calculate the means, variances and the relation coeffiient for these variables.

10-60. (*a*) Suppose that X and Y are independent random variables with

$$f(x) = 8x^{-3},\ x > 2,\ g(y) = 2y,\ 0 < y < 1$$

Evaluate the correlation coefficient between $X + Y$ and $X - Y$.

(*b*) Consider an experiment of tossing two tetrahedra. Let X be the number of the down ed face of the first tetrahedron and Y, the larger of the two numbers. Obtain the joint ribution of X and Y and hence $\rho(X, Y)$.

11·1. INTRODUCTION

The term *"regression"* literally means *"stepping back towards the average"*. It was first used by British biometrician Sir Francis Galton (1822 - 1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to *"regress"* or *"step back"* to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition. *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* and the variable which influences the values or is used for prediction is called *independent variable*. In regression analysis independent variable is also known as *regressor or predictor* or *explanatory variable* while the dependent variable is also known as *regressed* or *explained variable*.

11·2. LINEAR REGRESSION

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the *"curve of regression"*. If the curve is a straight line, it is called the line of regression and there is said to be *linear regression* between the variables, otherwise regression is said to be *curvilinear*.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of *"best fit"* and is obtained by the *principle of least squares*.

Let us suppose that in the bivariate distribution (x_i, y_i) ; $i = 1, 2, \ldots, n$; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be

$$Y = a + bx \qquad \ldots (11\cdot1)$$

The above equation (11·1) represents a family of straight lines for different values of the arbitrary constants *'a'* and *' b'*. The problem is to determine *'a'* and *'b'* so that the line (11·1) is the line of *'best fit'*.

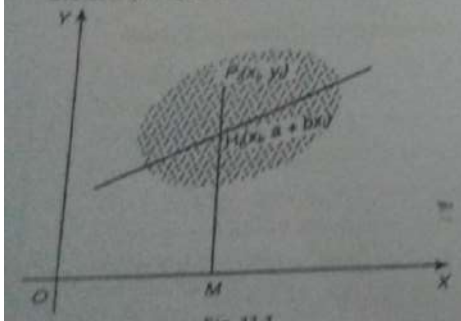


Fig. 11·1.

The term *'best fit'* is interpreted in accordance with Legender's principle of least squares which consists in minimising the sum of the squares of the deviations of the actual values of y from their estimated values as given by the line of best fit.

Let $P_i(x_i, y_i)$ be any general point in the scatter diagram (§ 10·2). Draw $P_iM \perp$ to x-axis meeting the line, (11·1) in H_i. Abscissa of H_i is x_i and since H_i lies on (11·1), its ordinate is $a + bx_i$. Hence the co-ordinates of H_i are $(x_i, a + bx_i)$.

$$P_iH_i = P_iM - H_iM = y_i - (a + bx_i),$$

is called the *error of estimate* or the *residual* for y_i.

According to the principle of least squares, we have to determine a and b so that

$$E = \sum_{i=1}^{n} P_iH_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

is minimum. From the principle of maxima and minima, the partial derivatives of E, w.r. to a and b should vanish separately, *i.e.*,

$$\frac{\partial E}{\partial a} = 0 \Rightarrow -2 \sum_{i=1}^{n} (y_i - a - bx_i) \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 \Rightarrow -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \qquad \ldots (11\cdot2) \quad \text{and} \quad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \qquad \ldots (11\cdot3)$$

Remark. Equations (11·2) and (11·3) are known as the *normal equations* for estimating a and b. All the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_iy_i$ can be obtained from the given set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$ and the equations (11·2) and (11·3) can be solved for a and b. With the values of a and b so obtained, equation (11·1) is the line of best fit to the given set of points $(x_i, y_i, i = 1, 2, \ldots, n$.

From (11·2) on dividing by n, we get $\quad \bar{y} = a + b\bar{x}$ $\qquad \ldots (11\cdot4)$

Thus, *the line of regression of Y on X passes through the point* $(\bar{x}, \bar{y})$.

Now $\quad \mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_iy_i = \mu_{11} + \bar{x}\bar{y} \qquad \ldots (11\cdot5)$

Also $\quad \sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \sigma_X^2 + \bar{x}^2 \qquad \ldots (11\cdot5a)$

Dividing (11·3) by n and using (11·5) and (11·5a), we get

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2)$$

Multiplying (11·4) by $\bar{x}$ and then subtracting from (11·6), we get

$$\mu_{11} = b\sigma_X^2 \Rightarrow b = \frac{\mu_{11}}{\sigma_X^2}$$

Since *'b'* is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is :

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x})$$

$$\Rightarrow \quad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

* The general problem in curve fitting is to find, if possible, an analytic expression of the form $y = f(x)$, for the functional relationship suggested by the given data. Fitting of such a set of numerical data is of considerable importance, theoretical as well as practical. From the theoretical point of view it is useful in the study of correlation and regression, *e.g.*, lines of regression can be obtained by the fitting of linear curves to the given bivariate distribution. In practical statistics it is frequently used to represent the relationship between two variables by simple algebraic expressions such as polynomials, exponential or logarithmic functions. Moreover, it may be used to estimate the values of one variable which would correspond to the specified values of the other variable.

Starting with the equation $X = A + BY$ and proceeding similarly or by simply interchanging the variables X and Y in (11·8) and (11·8*a*), the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y}) \quad \Rightarrow \quad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \qquad \ldots (11\cdot9), \ldots (11\cdot9a)$$

Aliter. The straight line defined by $Y = a + bX$ ...(*i*)

and satisfying the residual (least square) condition $S = E[(Y - a - bX)^2] = \text{Min.}$ ... (*ii*)

for variations in a and b, is called the line of regression of Y on X.

The necessary and sufficient conditions for a minima of S, subject to variations in a and b are :

$$(i)\ \frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0 \quad \ldots (*) \quad \text{and} \quad (ii)\ \Delta = \begin{vmatrix} \frac{\partial^2 S}{\partial a^2} & \frac{\partial^2 S}{\partial a \partial b} \\ \frac{\partial^2 S}{\partial b\, \partial a} & \frac{\partial^2 S}{\partial b^2} \end{vmatrix} > 0 \text{ and } \frac{\partial^2 S}{\partial a^2} > 0 \quad \ldots (**)$$

Using (*), we get

$$\frac{\partial S}{\partial a} = -2E[Y - a - bX] = 0 \qquad \ldots (iii)$$

$$\frac{\partial S}{\partial b} = -2E[X(Y - a - bX)] = 0 \qquad \ldots (iv)$$

$$\Rightarrow \quad E(Y) = a + bE(X) \quad \ldots (v) \quad \text{and} \quad E(XY) = aE(X) + bE(X^2) \quad \ldots (vi)$$

Equation (*v*) implies that the line (*i*) of regression of Y on X passes through the mean value $[E(X), E(Y)]$.

Multiplying (*v*) by $E(X)$ and subtracting from (*vi*), we get

$$E(XY) - E(X)\,E(Y) = b[E(X^2) - \{E(X)\}^2]$$

$$\Rightarrow \quad \text{Cov}(X, Y) = b\sigma_X^2 \quad \Rightarrow \quad b = \frac{\text{Cov}(X, Y)}{\sigma_X^2} = \frac{r\sigma_Y}{\sigma_X} \qquad \ldots (vii)$$

Subtracting (*v*) from (*i*) and using (*vii*), we obtain the equation of line of regression of Y on X as :

$$Y - E(Y) = \frac{\text{Cov}(X, Y)}{\sigma_X^2}[X - E(X)] \quad \Rightarrow \quad Y - E(Y) = \frac{r\sigma_Y}{\sigma_X}[X - E(X)]$$

Similarly, the straight line defined by $X = A + BY$ and satisfying the residual condition $E[X - A - BY]^2 = \text{Minimum}$, is called the line of regression of X on Y.

Remarks 1. We note that

$$\frac{\partial^2 S}{\partial a^2} = 2 > 0, \quad \frac{\partial^2 S}{\partial b^2} = 2E(X^2) \quad \text{and} \quad \frac{\partial^2 S}{\partial a \partial b} = 2E(X)$$

Substituting in (**), we have

$$\Delta = \frac{\partial^2 S}{\partial a^2}\cdot\frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b}\right)^2 = 4[E(X^2) - \{E(X)\}^2] = 4\sigma_X^2 > 0$$

Hence the solution of the least square equations (*iii*) and (*iv*), in fact provides a minima of

2. The regression equation (11·8*a*) implies that the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Similarly (11·9*a*) implies that the line of regression of X on Y also passes through the point $(\bar{x}, \bar{y})$. Hence both the lines of regression pass through the point $(\bar{x}, \bar{y})$. In other words, the mean values $(\bar{x}, \bar{y})$ can be obtained as the point of intersection of the two regression lines.

3. *Why two lines of Regression ?* There are always two lines of regression, one of Y on X and the other of X on Y. The line of regression of Y on X (11·8*a*) is used to estimate or predict the value of Y for any given value of X, *i.e.*, when Y is a dependent variable and X is an independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares. We can also obtain an estimate of X for any given value of Y by using equation (11·8*a*) but the estimate so obtained will not be best since (11·8*a*) is obtained on minimising the sum of the squares of errors of estimates in Y and not in X. Hence to estimate or predict X for any given value of Y, we use the regression equation of X on Y (11·9*a*) which is derived on minimising the sum of the squares of errors of estimates in X. Here X is a dependent variable and Y is an independent variable. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different. The regression equation of Y on X is obtained on minimising the sum of the squares of the errors parallel to the Y-axis while the regression equation of X on Y is obtained on minimising the sum of squares of the errors parallel to the X-axis.

In a particular case of perfect correlation, positive or negative, *i.e.*, $r = \pm 1$, the equation of line of regression of Y on X becomes :

$$Y - \bar{y} = \pm\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \quad \Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm\left(\frac{X - \bar{x}}{\sigma_X}\right) \qquad \ldots (11\cdot10)$$

Similarly, the equation of the line of regression of X on Y becomes :

$$Y - \bar{x} = \pm\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \quad \Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm\left(\frac{X - \bar{x}}{\sigma_X}\right)$$

which is same as (11·10).

Hence in case of perfect correlation, $(r = \pm 1)$, both the lines of regression coincide. Therefore, in general, we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

11·2·1. Regression Coefficients. 'b', the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X} \qquad \ldots (11\cdot11)$$

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by :

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y} \qquad \ldots (11\cdot1)$$

11·2·2. Properties of Regression Coefficients.

(*a*) *Correlation coefficient is the geometric mean between the regression coefficients.*

Proof. Multiplying (11·11) and (11·11*a*), we get

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2 \quad \Rightarrow \quad r = \pm\sqrt{b_{XY} \times b_{YX}}$$

Thus the correlation is the geometric mean of the regression coefficient : (113)

Remark. We have

$$r = \frac{\mu_{11}}{\sigma_X \cdot \sigma_Y}, b_{YX} = \frac{\mu_{11}}{\sigma_X^2} \text{ and } b_{XY} = \frac{\mu_{11}}{\sigma_Y^2}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term μ_{11}. Thus, if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

Hence, the sign to be taken before the square root in (11·12) is that of the regression coefficients.

(b) If one of the regression coefficients is greater than unity, the other must be less than unity.

Proof. Let one of the regression coefficients (say) b_{YX} be greater than unity, then we have to show that $b_{XY} < 1$.

Now $b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1$ ... (*) Also $r^2 \le 1 \Rightarrow b_{YX} \cdot b_{XY} \le 1$

Hence $b_{XY} \le \frac{1}{b_{YX}} < 1$ [From (*)]

(c) The modulus value of the arithmetic mean of the regression coefficients is not less than the modulus value of the correlation coefficient r.

Proof. We have to prove that $\left|\frac{1}{2}(b_{YX} + b_{XY})\right| > |r|$.

$$\Rightarrow \left|\frac{1}{2}\left(r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y}\right)\right| \ge |r| \Rightarrow \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \ge 2 \quad (\because |r| > 0)$$

$$\Rightarrow \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y > 0 \Rightarrow (\sigma_Y - \sigma_X)^2 > 0,$$

which is always true, since the square of a real quantity is ≥ 0.

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

Proof. Let $U = \frac{X-a}{h}, V = \frac{Y-b}{k} \Rightarrow X = a + hU, Y = b + kV,$

where a, b, h (> 0) and k (> 0) are constants.

Then $Cov(X, Y) = hk\, Cov(U, V)$, $\sigma_X^2 = h^2\sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$$\therefore \quad b_{YX} = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{hk\, Cov(U, V)}{h^2\sigma_U^2} = \frac{k}{h} \cdot \frac{Cov(U, V)}{\sigma_U^2} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that $b_{XY} = (h/k)\, b_{UV}$.

11·2·3. Angle Between Two Lines of Regression. Equations of the lines of regression of Y on X, and X on Y are :

$$Y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \text{ and } X - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \Rightarrow Y - \bar{y} = \frac{\sigma_Y}{r\sigma_X}(X - \bar{x})$$

Slopes of these lines are $r \cdot \frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the acute angle between the two lines of regression, then

$$\tan\theta = \left|\frac{r\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_Y}{r\sigma_X}}{1 + r\frac{\sigma_Y}{\sigma_X} \cdot \frac{\sigma_Y}{r\sigma_X}}\right| = \left|\frac{r^2 - 1}{r}\right|\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 - r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \quad (\because r^2 \le 1)$$

$$\therefore \quad \theta = \tan^{-1}\left\{\frac{1 - r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \quad \text{... (1)}$$

Case (i). $(r = 0)$. If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$. Thus if the two variables uncorrelated, the lines of regression become perpendicular to each other.

Case (ii). $(r = \pm 1)$. If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π. In this case the two line regression either coincide or they are parallel to each other. But since both line regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel. Hence in the cas perfect correlation, positive or negative, the two lines of regression coincide.

Remarks 1. *Whenever* two lines intersect, there are two angles between them, one angle and other obtuse angle. Further $\tan\theta > 0$ if $0 < \theta < \pi/2$, *i.e.*, θ is an acute angle and $\tan\theta < 0$ if $\pi/2 < \theta < \pi$, *i.e.*, θ is an obtuse angle and since $0 < r^2 < 1$, the acute angle (θ_1) and angle θ_2 between the two lines of regression are given by :

$$\theta_1 = \text{Acute angle} = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{1 - r^2}{|r|}\right\}, \text{ and } \theta_2 = \pi - \theta_1$$

2. When $r = 0$, *i.e.*, variables X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by :

[From (11·8a) and (11·9a)] $Y = \bar{Y}$ and $X = \bar{X}$

as shown in the adjoining diagram. Hence, in this case $(r = 0)$, the lines of regression are perpendicular to each other and are parallel to X-axis and Y-axis respectively.

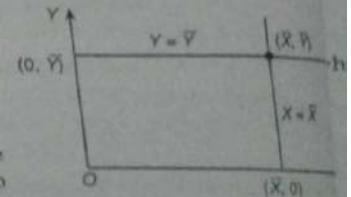


Fig. 11·2

3. The fact that if $r = 0$ (variables uncorrelated), the two lines of regressi perpendicular to each and if $r = \pm 1$, $\theta = 0$, *i.e.*, the two lines coincide, leads us to the con that for higher degree of correlation between the variables, the angle between the smaller, *i.e.*, the two lines of regression are nearer to each other. On the other hand, if of regression make a larger angle, they indicate a poor degree of correlation betw variables and ultimately for $\theta = \pi/2$, $r = 0$, *i.e.*, the lines become perpendicular if no co exists between the variables. Thus by plotting the lines of regression on a graph paper have an approximate idea about the degree of correlation between the two variab study. Consider the following illustrations :

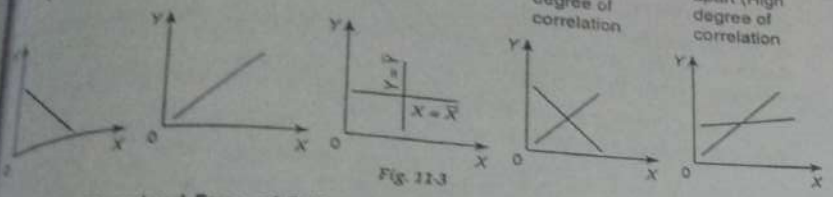


Fig. 11.3

11·2·4. Standard Error of Estimate or Residual Variance. The equation of the line of regression of Y on X is :

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \quad \Rightarrow \quad \frac{Y-\bar{Y}}{\sigma_Y} = r\frac{X-\bar{X}}{\sigma_X}$$

The residual variance S_Y^2 is the expected value of the squares of deviations of the observed values of Y from the expected values as given by the line of regression of Y on X. Thus

$$S_Y^2 = E[Y - \hat{Y}]^2 = E\left[Y - \{\bar{Y} + (r\sigma_Y(X-\bar{X})/\sigma_X)\}\right]^2.$$

$$= \sigma_Y^2 E\left[\frac{Y-\bar{Y}}{\sigma_Y} - r\left(\frac{X-\bar{X}}{\sigma_X}\right)\right]^2 = \sigma_Y^2 E(Y^* - rX^*)^2,$$

where X^* and Y^* are standardised variates so that

$$E(X^{*2}) = 1 = E(Y^{*2}) \text{ and } E(X^*Y^*) = r.$$

$$\therefore \quad s_Y^2 = \sigma_Y^2\left[E(Y^{*2}) + r^2 E(X^{*2}) - 2r E(X^*Y^*)\right] = \sigma_Y^2(1-r^2)$$

$$\Rightarrow \quad s_Y = \sigma_Y(1-r^2)^{1/2} \qquad \ldots (11\cdot13a)$$

Similarly, the standard error of estimate of X is given by :

$$s_X = \sigma_X(1-r^2)^{1/2} \qquad \ldots (11\cdot13b)$$

Remarks 1. Since s_X^2 or $s_Y^2 \geq 0$, it follows that $(1-r^2) \geq 0 \quad \Rightarrow \quad |r| \leq 1$.

Hence $\quad -1 \leq r(X, Y) \leq 1.$

2. If $r = \pm 1$, $s_X = s_Y = 0$ so that each deviation is zero, and the two lines of regression are coincident.

3. Since, as $r^2 \to 1$, s_X and $s_Y \to 0$, it follows that departure of the value r^2 from unity indicates the departure of the relationship between the variables X and Y from linearity.

4. From the definition of linear regression, the minima condition implies that s^2_Y or s^2_X is the minimum variance.

11·2·5. Correlation Coefficient Between Observed and Estimated Values. Here we will find the correlation between Y and $\hat{Y} = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$,

where $\hat{Y}$ is the estimated value of Y as given by the line of regression of Y on X.

We have : $$r(Y, \hat{Y}) = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_Y \sigma_{\hat{Y}}}$$

We have $E(\hat{Y}) = E\left[\bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X-\bar{X})\right] = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}E(X-\bar{X}) = \bar{Y}$

$$\therefore \quad \text{Var}(\hat{Y}) = E(\hat{Y} - E(\hat{Y}))^2 = E\left[r\frac{\sigma_Y}{\sigma_X}(X-\bar{X})\right]^2 = r^2\sigma_Y^2$$

$$\Rightarrow \quad \sigma_{\hat{Y}} = r\sigma_Y$$

Also $\text{Cov}(Y, \hat{Y}) = E\left[(Y - E(Y))(\hat{Y} - E(\hat{Y}))\right]$

$$= E\left[b\{(X - E(X))\}\left\{r\frac{\sigma_Y}{\sigma_X}(X-\bar{X})\right\}\right]$$

$$= b\,r\frac{\sigma_Y}{\sigma_X}E\left[(X - E(X))^2\right] = \left(r\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 = r^2\sigma_Y^2 \left[\because b = b_{YX} = r\frac{\sigma_Y}{\sigma_X}\right]$$

$$\therefore \quad r(Y, \hat{Y}) = \frac{r^2\sigma_Y^2}{\sigma_Y r\sigma_Y} = r = r(X, Y)$$

Hence the correlation coefficient between observed and estimated value of Y is the same as the correlation coefficient between X and Y.

***Example* 11·1.** *Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for $Y = 70$.*

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution. Let $U = X - 68$ and $V = Y - 69$, then

$\bar{U} = 0, \bar{V} = 0, \sigma_U^2 = 4{\cdot}5, \sigma_V^2 = 5{\cdot}5$, $\text{Cov}(U, V) = 3$ and $r(U, V) = 0{\cdot}6$ (c.f. page 10·

Since correlation coefficient is independent of change of origin, we get

$$r = r(X, Y) = r(U, V) = 0{\cdot}6$$

We know that if $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, then

$$\bar{X} = a + h\bar{U}, \quad \bar{Y} = b + k\bar{V}, \quad \sigma_X = h\sigma_U \text{ and } \sigma_Y = k\sigma_Y$$

Here we are given : $h = k = 1$, $a = 68$ and $b = 69$.

Thus $\quad \bar{X} = 68 + 0 = 68, \bar{Y} = 69 + 0 = 69$

$$\sigma_X = \sigma_U = \sqrt{4{\cdot}5} = 2{\cdot}12 \text{ and } \sigma_Y = \sigma_V = \sqrt{5{\cdot}5} = 2{\cdot}35$$

Equation of line of regression of Y on X is : $Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$

i.e., $\quad Y = 69 + 0{\cdot}6 \times \frac{2{\cdot}35}{2{\cdot}12}(X - 68) \quad \Rightarrow \quad Y = 0{\cdot}665X + 23{\cdot}78$

Equation of line of regression of X on Y is :

$$X - \bar{X} = r\,\frac{\sigma_x}{\sigma_y}\,(Y - \bar{Y})$$

$$\Rightarrow \qquad X = 68 + 0{\cdot}6 \times \frac{2{\cdot}12}{2{\cdot}35}(Y - 69) \quad \Rightarrow \quad X = 0{\cdot}54Y + 30{\cdot}74$$

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by : $\hat{X} = 0{\cdot}54 \times 70 + 30{\cdot}74 = 68{\cdot}54$,

where $\hat{X}$ is estimate of X.

Example 11·2. *In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are legible :*

Variance of $X = 9$. Regression equations : $8X - 10Y + 66 = 0$, $40X - 18Y = 214$.

What are : (i) the mean values X and Y, (ii) the correlation coefficient between X and Y, and (iii) the standard deviation of Y ?

Solution. (i) Since both lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have

$8\bar{X} - 10\bar{Y} + 66 = 0$, and $40\bar{X} - 18\bar{Y} = 214$. Solving, we get $\bar{X} = 13$, $\bar{Y} = 17$.

(ii) Let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the lines of regression of Y on X and X on Y respectively. These equations can be put in the form :

$$Y = \frac{8}{10}X + \frac{66}{10} \qquad \text{and} \qquad X = \frac{18}{40}Y + \frac{214}{40}$$

$\therefore$ b_{YX} = Regression coefficient of Y on $X = \frac{8}{10} = \frac{4}{5}$

and b_{XY} = Regression coefficient of X on $Y = \frac{18}{40} = \frac{9}{20}$

Hence $r^2 = b_{YX} \cdot b_{XY} = \frac{4}{5} \cdot \frac{9}{20} = \frac{9}{25}$ $\therefore$ $r = \pm\frac{3}{5} = \pm 0{\cdot}6$

But since both the regression coefficients are positive, we take $r = +0{\cdot}6$.

(iii) We have $b_{YX} = r\,\frac{\sigma_y}{\sigma_x} \Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_y}{3}$

Hence $\sigma_y = 4$.

Remarks 1. It can be verified that the values of $\bar{X} = 13$ and $\bar{Y} = 17$ as obtained in part (i) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that $8X - 10Y + 66 = 0$, is the equation of the line of regression of X on Y and $40X - 18Y = 214$ is the equation of line of regression of Y on X, then we get respectively :

$$8X = 10Y - 66 \qquad \text{and} \qquad 18Y = 40X - 214$$

$$\Rightarrow \quad X = \frac{10}{8}Y - \frac{66}{8} \qquad \text{and} \qquad Y = \frac{40}{18}X - \frac{214}{18}$$

$$\Rightarrow \quad b_{XY} = \frac{10}{8} \qquad \text{and} \qquad b_{YX} = \frac{40}{18}$$

$$\therefore \quad r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2{\cdot}78$$

But since r^2 always lies between 0 and 1, our supposition is wrong.

Example 11·3. *Find the most likely price in Mumbai corresponding to the price of Rs. at Kolkata from the following :*

	Kolkata	Mumbai
Average price	65	67
Standard deviation	2·5	3·5

Correlation coefficient between the prices of commodities in the two cities is 0·8.

Solution. Let the prices (in Rupees) in Kalkata and Mumbai be denoted by X, Y respectively. Then we are given :

$\bar{X} = 65$, $\bar{Y} = 67$, $\sigma_X = 2{\cdot}5$, $\sigma_Y = 3{\cdot}5$ and $r = r(X, Y) = 0{\cdot}8$. We want Y for $X =$

Line of regression of Y on X is :

$$Y - \bar{Y} = r\,\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \quad \Rightarrow \quad Y = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(X - 65)$$

When $X = 70$, $\hat{Y} = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(70 - 65) = 72{\cdot}6$

Hence, the most likely price in Mumbai corresponding to the price of Rs. Kolkata is Rs. 72·60.

Example 11·4. *Can $Y = 5 + 2{\cdot}8X$ and $X = 3 - 0{\cdot}5Y$ be the estimated regression equations of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.*

Solution. Line of regression of Y on X is : $Y = 5 + 2{\cdot}8X \Rightarrow b_{YX} = 2{\cdot}8$

Line of regression of X on Y is : $X = 3 - 0{\cdot}5Y \Rightarrow b_{XY} = -0{\cdot}5$

This is not possible, since each of the regression coefficients b_{YX} and b_{XY} must the same sign, which is same as that of Cov (X, Y). If Cov (X, Y) is positive, then the regression coefficients are positive and if Cov (X, Y) is negative, then both regression coefficients are negative. Hence (*) and (**) cannot be the estimated regression equations of Y on X and X on Y respectively.

Example 11·5. *By minimising $\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2$ for variations in α and p, show that there are two straight lines passing through the mean of the distribution for which the sum of squares of normal deviations has an extreme value. Prove also that their inclinations are given by :*

$$\tan 2\alpha = \frac{2\mu_{11}}{\sigma_x^2 - \sigma_y^2}.$$

Solution. We have to minimize : $S = \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2$

for variations in α and p.

Equating to zero, the partial derivatives of (*) w.r. to α and p, we have

$$\frac{\partial S}{\partial \alpha} = 0 = 2\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(-x_i \sin\alpha + y_i \cos\alpha)$$

$$\frac{\partial S}{\partial p} = 0 = -2\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)$$

Equation (*iii*) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) = 0 \quad \Rightarrow \quad \bar{x}\cos\alpha + \bar{y}\sin\alpha - p = 0 \qquad \ldots (iv)$$

From equation (*ii*), we get a quadratic equation which shows that there are two straight lines for extreme values of E.

From equation (*iv*), it becomes clear that both the straight lines pass through the point $(\bar{x}, \bar{y})$.

Again equation (*ii*) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(y_i \cos\alpha - x_i \sin\alpha) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \left\{ f_i \left[\cos\alpha (x_i - \bar{x}) + \sin\alpha (y_i - \bar{y})\right] (y_i \cos\alpha - x_i \sin\alpha) \right. = 0 \quad \text{[Using (iv)]}$$

$$\Rightarrow \quad \cos^2\alpha \sum_{i=1}^{n} f_i y_i (x_i - \bar{x}) - \sin\alpha\cos\alpha \sum_{i=1}^{n} f_i x_i (x_i - \bar{x})$$

$$+ \sin\alpha\cos\alpha \sum_{i=1}^{n} f_i y_i (y_i - \bar{y}) - \sin^2\alpha \sum_{i=1}^{n} f_i x_i (y_i - \bar{y}) = 0 \qquad \ldots (v)$$

We have $\mu_{11} = \frac{1}{N} \sum_i f_i (x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{N} \sum_i f_i x_i (y_i - \bar{y}) - \bar{x} \cdot \frac{1}{N} \sum_i f_i (y_i - \bar{y}) = \frac{1}{N} \sum f_i x_i (y_i - \bar{y})$$

Similarly, $\mu_{11} = \frac{1}{N} \sum_i f_i y_i (x_i - \bar{x})$

$$\sigma_X^2 = \frac{1}{N} \sum_i f_i x_i (x_i - \bar{x}) \quad \text{and} \quad \sigma_Y^2 = \frac{1}{N} \sum_i f_i y_i (y_i - \bar{y})$$

Substituting these values in (*v*), we get the required result.

11·3. CURVILINEAR REGRESSION

One of the criteria set forth previously in our regression studies has been that the variables X and Y are related linearly. In many cases, this assumption may not be valid. For example, in a scatter diagram showing the relationship between X, the number of overtime hours, versus Y, the resulting number of additional units produced for a manufacturing firm, it appears by inspection that a curvlinear regression may explain more of the variability of Y than does a linear line. For this reason, we shall fit the regression line using the following formula :

$$Y = a + b_1 X + b_2 X^2 \qquad \ldots (11\cdot14)$$

Using the principle of least squares, we have to determine the constants a, b_1 and b_2 so that : $E = \sum_{i=1}^{n} (y_i - a - b_1 x_i - b_2 x_i^2)^2$ is minimum.

Equating to zero the partial derivatives of E with respect to a, b and c separately, we get the normal equations for estimating a, B and C as

$$\left.\begin{aligned} \frac{\partial E}{\partial a} &= 0 = -2\sum_i (y_i - a - b_1 x_i - b_2 x_i^2) \\ \frac{\partial E}{\partial b_1} &= 0 = -2\sum_i x_i (y_i - a - b_1 x_i - b_2 x_i^2) \\ \frac{\partial E}{\partial b_2} &= 0 = -2\sum_i x_i^2 (y_i - a - b_1 x_i - b_2 x_i^2) \end{aligned}\right\} \qquad \ldots (11\cdot15)$$

$$\Rightarrow \quad \left\{\begin{aligned} \sum y_i &= na + b_1 \sum x_i + b_1 \sum x_i^2 \\ \sum x_i y_i &= a\sum x_i + b_1 \sum x_i^2 + b_1 \sum x_i^3 \\ \sum x_i^2 y_i &= a\sum x_i^2 + b_1 \sum x_i^3 + b_1 \sum x_i^4 , \end{aligned}\right. \qquad \ldots (11\cdot15a)$$

summation taken over i from 1 to n.

For given set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$, equations (11·15a) can be solved for a, b_1 and b_2, and with these values of a, b_1 and b_2 (11·14) is the parabola of best fit.

Remarks **1 .** If $Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_k X^k$ (11·16)

is the kth degree polynomial of best fit to the set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$ the constants $a_0, a_1, a_2, \ldots, a_k$ are to be obtained so that

$$E = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k)^2$$

is minimum. Thus, the normal equations for estimating $a_0, a_1, \ldots, a_k$ are obtained on equating to zero the partial derivatives of E *w.r.to.* $a_0, a_1, \ldots, a_k$ separately, *i.e.*,

$$\left.\begin{aligned} \frac{\partial E}{\partial a_0} &= 0 = -2\sum (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k) \\ \frac{\partial E}{\partial a_1} &= 0 = -2\sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k) \\ &\vdots \\ \frac{\partial E}{\partial a_k} &= 0 = -2\sum x_i^k (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k) \end{aligned}\right\} \qquad \ldots (11\cdot17)$$

$$\Rightarrow \quad \left.\begin{aligned} \sum y_i &= na_0 + a_1 \sum x_i + a_2 \sum x_i^2 + \ldots + a_k \sum x_i^k \\ \sum x_i y_i &= a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 + \ldots + a_k \sum x_i^{k+1} \\ &\vdots \\ \sum x_i^k y_i &= a_0 \sum x_i^k + a_1 \sum x_i^{k+1} + a_2 \sum x_i^{k+2} + \ldots + a_k \sum x_i^{2k} \end{aligned}\right\} \qquad \ldots (11\cdot17a)$$

summation extended over i from 1 to n. These are $(k + 1)$ equations in $(k + 1)$ unknowns $a_0, a_1, a_2, \ldots, a_k$ and can be solved with the help of algebra.

2. Depending on the pattern suggested by the scatter diagram, there are many non-linear equations from which we can select a regression line. In such cases' the original data which is not in a linear form can be reduced to linear form by simple transformation of variables. This will be illustrated by considering the following curves. ... (11·18)

(*a*) Fitting of a Power Curve $Y = aX^b$ to a set of n points.

Taking logarithm of each side, we get

$$\log Y = \log a + b \log X \quad \Rightarrow \quad U = A + bV,$$

where $U = \log Y$, $A = \log a$ and $V = \log X$.

This is a linear equation in V and U.

Normal equations for estimating A and B are : ... (11·19)

$$\Sigma U = nA + b\Sigma V \quad \text{and} \quad \Sigma UV = A\Sigma V + b\Sigma V^2$$

$$P(X+Y \le 16) = P\left(Z \le \frac{1}{\sqrt{26}}\right) = \Phi\left(1/\sqrt{26}\right),$$

where $\Phi(z) = P(Z \le z)$, is the distribution function of standard normal variate.

Remark. $P(X+Y \le 16) = P\left(Z \le \frac{1}{5\sqrt{0.99}}\right) = P(Z \le 0.196)$

$= 0.5 + P(0 \le Z \le 0.196) = 0.5 + 0.0793$ (approx.)

$= 0.5793$.

12·4. MULTIPLE AND PARTIAL CORRELATION

When the values of one variable are associated with or influenced by other variables, *e.g.*, the age of husband and wife, the height of father and son, the supply and demand of a commodity and so on, Karl Pearson's coefficient of correlation can be used as a measure of linear relationship between them. But sometimes there is interrelation between many variables and the value of one variable may be influenced by many others, *e.g.*, the yield of crop per acre, say, (X_1) depends upon quality of seed (X_2), fertility of soil (X_3), fertilizer used (X_4), irrigation facilities (X_5), weather conditions (X_6) and so on. Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in that group, our study is that of *multiple correlation and multiple regression*.

Suppose in a trivariate or multi-variate distribution we are interested in the relationship between two variables only. There are two alternatives, *viz.*, (*i*) we consider only those two members of the observed data in which the other members have specified values or (*ii*) we may eliminate mathematically the effect of other variates on two variates. The first method has the disadvantage that it limits the size of the data and also it will be applicable to only the data in which the other variates have assigned values. In the second method it may not be possible to eliminate the entire influence of the variates but the linear effect can be easily eliminated. The correlation and regression between only two variates eliminating the linear effect of other variates in them is called the *partial correlation and partial regression*.

12·4·1. Yule's Notation. Let us consider a distribution involving three random variables X_1, X_2 and X_3. Then the equation of the plane of regression of X_1 on X_2 and X_3 is :

$$X_1 = a + b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3 \qquad \ldots (12\cdot8)$$

Without loss of generality we can assume that the variables X_1, X_2 and X_3 have been measured from their respective means, so that $E(X_1) = E(X_2) = E(X_3) = 0$ ... (12·8*a*)

On taking expectation of both sides in (12·8), we get $a = 0$.

Thus the plane of regression of X_1 on X_2 and X_3 becomes

$$X_1 = b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3 \qquad \ldots (12\cdot8b)$$

The coefficients $b_{12\cdot3}$ and $b_{13\cdot2}$ are known as the *partial regression coefficients* of X_2 and of X_1 on X_3 respectively.

$$e_{1\cdot23} = b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3$$

the estimate of X_1 as given by the plane of regression (12·8*b*) and the quantity

$$X_{1\cdot23} = X_1 - b_{12\cdot3}\,X_2 - b_{13\cdot2}\,X_3 = X_1 - e_{1\cdot23}, \qquad \ldots (12\cdot8c)$$

the *error of estimate or residual*.

In the general case of n variable $X_1, X_2, \ldots, X_n$ the equation of the regression of X_1 on $X_2, X_3 \ldots X_n$ becomes

$$X_1 = b_{12\cdot34\cdots n}\,X_2 + b_{13\cdot24\cdots n}\,X_3 + \ldots + b_{1n\cdot23\cdots(n-1)}\,X_n$$

The error of estimate or residual is given by :

$$X_{1\cdot23\cdots n} = X_1 - (b_{12\cdot34\cdots n}\,X_2 + b_{13\cdot24\cdots n}\,X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}\,X_n)$$

The notations used here are due to yule. The subscripts before the dot are known as *primary subscripts* and those after the dot are called *secondary subscripts*. The order of a regression coefficient is determined by the number of secondary subscripts, *i.e.*, $b_{12\cdot3}, b_{12\cdot34}, \ldots, b_{12\cdot34\cdots n}$ are the regression coefficients of order 1, 2, ... respectively. Thus in general, a regression coefficient with p-secondary subscripts will be called a regression co-efficient of order 'p'. It may be pointed out that the order in which the secondary subscripts are written is immaterial but the order of the primary subscripts is important, *e.g.*, in $b_{12\cdot34\cdots n}$, X_2 is independent while X_1 is dependent variable but in $b_{21\cdot34\cdots n}$, X_1 is independent while X_2 is dependent variable. Thus two primary subscripts, former refers to dependent variable and the latter independent variable.

The order of a residual is also determined by the number of secondary subscripts in it, *e.g.*, $X_{1\cdot23}, X_{1\cdot234}, \ldots, X_{1\cdot23\cdots n}$ are the residuals of order 2, 3, ..., $(n-1)$ respectively.

Remark. In the following sequences we shall assume that the variables under consideration have been measured from their respective means so that $E(X_i) = 0$; $i = 1, 2, \ldots, n$.

12·5. PLANE OF REGRESSION

The equation of the plane of regression of X_1 on X_2 and X_3 is :

$$X_1 = b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3$$

The constants b' in (12·9) are determined by the principle of least squares, minimising the sum of the squares of the residuals, *viz.*,

$$S = \sum X_{1\cdot23}^2 = \sum (X_1 - b_{12\cdot3}\,X_2 - b_{13\cdot2}\,X_3)^2,$$

the summation being extended to the given values (N in number) of the variables.

Here we make N observations on each of the variables X_1, X_2 and X_3.

The normal equations for estimating $b_{12\cdot3}$ and $b_{13\cdot2}$ are :

$$\left.\begin{aligned} \frac{\partial S}{\partial b_{12\cdot3}} &= 0 = -2\sum X_2(X_1 - b_{12\cdot3}\,X_2 - b_{13\cdot2}\,X_3) \\ \frac{\partial S}{\partial b_{13\cdot2}} &= 0 = -2\sum X_3(X_1 - b_{12\cdot3}\,X_2 - b_{13\cdot2}\,X_3) \end{aligned}\right\}$$

i.e., $\sum X_2\,X_{1\cdot23} = 0$ and $\sum X_3\,X_{1\cdot23} = 0$

$$\Rightarrow \left.\begin{aligned} \sum X_1X_2 - b_{12\cdot3}\sum X_2^2 - b_{13\cdot2}\sum X_2X_3 &= 0 \\ \sum X_1X_3 - b_{12\cdot3}\sum X_2X_3 - b_{13\cdot2}\sum X_3^2 &= 0 \end{aligned}\right\}$$

Since X_i's are measured from their respective means, we have

$$\left.\begin{aligned} \sigma_i^2 = \frac{1}{N}\sum X_i^2,\ \mathrm{Cov}(X_i, X_j) &= \frac{1}{N}\sum X_iX_j \\ \text{and} \qquad r_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_i\sigma_j} &= \frac{\sum X_iX_j}{N\sigma_i\sigma_j} \end{aligned}\right\}$$

Hence from (12·10*b*), we get

$$\left.\begin{aligned} r_{12}\,\sigma_1\sigma_2 - b_{12\cdot3}\,\sigma_2^2 - b_{13\cdot2}\,r_{23}\,\sigma_2\sigma_3 = 0 \\ r_{13}\,\sigma_1\sigma_3 - b_{12\cdot3}\,r_{23}\,\sigma_2\,\sigma_3 - b_{13\cdot2}\,\sigma_3^2 = 0 \end{aligned}\right\} \quad \ldots (12\cdot10d)$$

Solving the equations (12·10*d*) for $b_{12\cdot3}$ and $b_{13\cdot2}$, we get

$$b_{12\cdot3} = \frac{\begin{vmatrix} r_{12}\sigma_1 & r_{23}\,\sigma_3 \\ r_{13}\,\sigma_1 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} \sigma_2 & r_{23}\,\sigma_3 \\ r_{23}\,\sigma_2 & \sigma_3 \end{vmatrix}} = \frac{\sigma_1}{\sigma_2}\cdot\frac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \quad \ldots (12\cdot11)$$

Similarly, we will get $$b_{13\cdot2} = \frac{\sigma_1}{\sigma_3}\cdot\frac{\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \quad \ldots (12\cdot11a)$$

If we write $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} \quad \ldots (12\cdot12)$$

and ω_{ij} is the cofactor of the element in the ith row and jth column of ω, we have from (12·11) and (12·11*a*) :

$$b_{12\cdot3} = -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{13\cdot2} = -\frac{\sigma_1}{\sigma_3}\cdot\frac{\omega_{13}}{\omega_{11}} \quad \ldots (12\cdot13)$$

Substituting these values in (12·9), we get the required equation of the plane of regression of X_1 on X_2 and X_3 as :

$$X_1 = -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}\cdot X_2 - \frac{\sigma_1}{\sigma_3}\cdot\frac{\omega_{13}}{\omega_{11}}X_3 \quad \Rightarrow \quad \frac{X_1}{\sigma_1}\,\omega_{11} + \frac{X_2}{\sigma_2}\cdot\omega_{12} + \frac{X_3}{\sigma_3}\,\omega_{13} = 0 \ldots (12\cdot14)$$

Alter. Eliminating the coefficient $b_{12\cdot3}$ and $b_{13\cdot2}$ in (12·9) and (12·10*d*), the required equation of the plane of regression of X_1 on X_2 and X_3 becomes :

$$\begin{vmatrix} X_1 & X_2 & X_3 \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 \\ r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{vmatrix} = 0$$

Dividing C_1, C_2 and C_3 by σ_1, σ_2 and σ_3 respectively and also R_2 and R_3 by σ_2 and σ_3 respectively, we get

$$\begin{vmatrix} \frac{X_1}{\sigma_1} & \frac{X_2}{\sigma_2} & \frac{X_3}{\sigma_3} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0 \quad \Rightarrow \quad \frac{X_1}{\sigma_1}\,\omega_{11} + \frac{X_2}{\sigma_2}\,\omega_{12} + \frac{X_3}{\sigma_3}\,\omega_{13} = 0,$$

where ω_{ij} is defined in (12·12).

12·5·1. Generalisation. In general, the equation of the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$ is :

$$X_1 = b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}X_n \quad \ldots (12\cdot15)$$

The sum of the squares of residuals is given by

$$S = \Sigma X^2_{1\cdot23\ldots n} = \Sigma\left[X_1 - b_{12\cdot34\ldots n}X_2 - b_{13\cdot24\ldots n}X_3 - \ldots - b_{1n\cdot23\ldots(n-1)}X_n\right]^2$$

Using the principle of least squares, the normal equations for estimating the $(n-1)$, b's are :

$$\left.\begin{aligned} \frac{\partial S}{\partial b_{12\cdot34\ldots n}} &= 0 = -2\Sigma\,X_2(X_1 - b_{12\cdot34\ldots n}X_2 - b_{13\cdot24\ldots n}X_3 - \ldots - b_{1n\cdot23\ldots(n-1)}X_n) \\ \frac{\partial S}{\partial b_{13\cdot24\ldots n}} &= 0 = -2\Sigma\,X_3(X_1 - b_{12\cdot34\ldots n}X_2 - b_{13\cdot24\ldots n}X_3 - \ldots - b_{1n\cdot23\ldots(n-1)}X_n) \\ &\vdots \\ \frac{\partial S}{\partial b_{1n\cdot23\ldots(n-1)}} &= 0 = -2\Sigma\,X_n(X_1 - b_{12\cdot34\ldots n}X_2 - b_{13\cdot24\ldots n}X_3 - \ldots - b_{1n\cdot23\ldots(n-1)}X_n) \end{aligned}\right\} \quad \ldots (12\cdot16)$$

$$i.e., \quad \Sigma X_i\,X_{1\cdot23\ldots n} = 0, \ (i = 2, 3, \ldots, n) \quad \ldots (12\cdot16a)$$

which on simplification after using (12·10*c*), give

$$\left.\begin{aligned} r_{12}\sigma_1\sigma_2 &= b_{12\cdot34\ldots n}\sigma_2^2 + b_{13\cdot24\ldots n}r_{23}\sigma_2\sigma_3 + \ldots + b_{1n\cdot23\ldots(n-1)}r_{2n}\sigma_2\sigma_n \\ r_{13}\sigma_1\sigma_3 &= b_{12\cdot34\ldots n}r_{23}\sigma_2\sigma_3 + b_{13\cdot24\ldots n}\sigma_3^2 + \ldots + b_{1n\cdot23\ldots(n-1)}r_{3n}\sigma_3\sigma_n \\ &\vdots \\ r_{1n}\sigma_1\sigma_n &= b_{12\cdot34\ldots n}r_{2n}\sigma_2\sigma_n + b_{13\cdot24\ldots n}r_{3n}\sigma_3\sigma_n + \ldots + b_{1n\cdot23\ldots(n-1)}\sigma_n^2 \end{aligned}\right\} \quad \ldots (12\cdot16b)$$

Hence the eliminant of b's between (12·15) and (12·16*b*) is :

$$\begin{vmatrix} X_1 & X_2 & X_3 & \ldots & X_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 & \ldots & r_{2n}\sigma_2\sigma_n \\ r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 & \ldots & r_{3n}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & r_{3n}\sigma_3\sigma_n & \ldots & \sigma_n^2 \end{vmatrix} = 0$$

Dividing $C_1, C_2, \ldots, C_n$ by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively and $R_2, R_3, \ldots, R_n$ by $\sigma_2, \sigma_3, \ldots, \sigma_n$ respectively, we get

$$\begin{vmatrix} \frac{X_1}{\sigma_1} & \frac{X_2}{\sigma_2} & \frac{X_3}{\sigma_3} & \ldots & \frac{X_n}{\sigma_n} \\ r_{12} & 1 & r_{32} & \ldots & r_{2n} \\ r_{13} & r_{23} & 1 & \ldots & r_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1n} & r_{2n} & r_{3n} & & 1 \end{vmatrix} = 0$$

If we write

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1n} \\ r_{21} & 1 & r_{23} & \ldots & r_{2n} \\ r_{31} & r_{32} & 1 & \ldots & r_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \ldots & 1 \end{vmatrix}$$

and ω_{ij} is the cofactor of the element in the ith row and jth column of ω, we get from

$$\frac{X_1}{\sigma_1}\omega_{11} + \frac{X_2}{\sigma_2}\omega_{12} + \frac{X_3}{\sigma_3}\omega_{13} + \ldots + \frac{X_n}{\sigma_n}\omega_{1n} = 0 \quad \ldots (12\cdot19)$$

(12·17) :

as the required equation of the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$.

Equation (12·19) can be re-written as :

$$X_1 = -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}X_2 - \frac{\sigma_1}{\sigma_3}\cdot\frac{\omega_{13}}{\omega_{11}}X_3 - \ldots - \frac{\sigma_1}{\sigma_n}\cdot\frac{\omega_{1n}}{\omega_{11}}X_n \quad \ldots (12\cdot19a)$$

Comparing (12·19a) with (12·15), we get

$$\left.\begin{aligned} b_{12\cdot34\ldots n} &= -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \\ b_{13\cdot24\ldots n} &= -\frac{\sigma_1}{\sigma_3}\cdot\frac{\omega_{13}}{\omega_{11}} \\ &\vdots \\ b_{1n\cdot23\ldots(n-1)} &= -\frac{\sigma_1}{\sigma_n}\cdot\frac{\omega_{1n}}{\omega_{11}} \end{aligned}\right\} \quad \ldots (12\cdot20)$$

Remarks 1. From the symmetry of the result obtained in (12·20) the equation of the plane of regression of X_i, (say), on the remaining variables X_j $(j \neq i = 1, 2, \ldots, n)$, is given by :

$$\frac{X_1}{\sigma_1}\omega_{i1} + \frac{X_2}{\sigma_2}\omega_{i2} + \ldots + \frac{X_i}{\sigma_i}\omega_{ii} + \ldots + \frac{X_n}{\sigma_n}\omega_{in} = 0\,;\, i = 1, 2, \ldots, n \quad \ldots (12\cdot21)$$

2. We have

$$b_{12\cdot34\ldots n} = -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{21\cdot34\ldots n} = -\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}}$$

Since each of σ_1, σ_2, ω_{11} and ω_{22} is non-negative and $\omega_{12} = \omega_{21}$, [*c.f.* Remarks 3 and 4 to §12·7]. the sign of each regression coefficient $b_{12\cdot34\ldots n}$ and $b_{21\cdot34\ldots n}$ depends on ω_{12}.

12·6. PROPERTIES OF RESIDUALS

Property 1. *The sum of the product of any residual of order zero with any other residual of higher order is zero, provided the subscript of the former occurs among the secondary subscripts of the latter.*

The normal equations for estimating b's in trivariate and n-variate distributions, as obtained in equations (12·10a) and (12·16a), are :

$$\Sigma X_2 X_{1\cdot23} = 0, \Sigma X_3 X_{1\cdot23} = 0 \quad \text{and} \quad \Sigma X_i X_{1\cdot23\ldots n} = 0\,;\, i = 2, 3, \ldots, n$$

respectively. Here X_i, $(i = 1, 2, 3, \ldots, n)$ can be regarded as a residual of order zero. Hence the result.

Property 2. *The sum of the product of any two residuals in which all the secondary subscripts of the first occur among the secondary subscripts of the second is unaltered if we omit any or all of the secondary subscripts of the first. Conversely, the product sum of any residual of order 'p' with a residual of order p + q, the 'p' subscripts being the same in each case is unaltered by adding to the secondary subscripts of the former any or all the 'q' additional subscripts of the latter.*

Let us consider :

$$\begin{aligned} \Sigma X_{1\cdot2} X_{1\cdot23} &= \Sigma (X_1 - b_{12}X_2)X_{1\cdot23} = \Sigma X_1 X_{1\cdot23} - b_{12}\Sigma X_2 X_{1\cdot23} \\ &= \Sigma X_1 X_{1\cdot23} \qquad \text{(Using Property 1)} \end{aligned}$$

$$\begin{aligned} \text{Also } \Sigma X_{1\cdot23}^2 &= \Sigma X_{1\cdot23} X_{1\cdot23} = \Sigma (X_1 - b_{12\cdot3}X_2 - b_{13\cdot2}X_3) X_{1\cdot23} \\ &= \Sigma X_1 X_{1\cdot23} - b_{12\cdot3}\Sigma X_2 X_{1\cdot23} - b_{13\cdot2}\Sigma X_3 X_{1\cdot23} \\ &= \Sigma X_1 X_{1\cdot23} \qquad \text{(Using Prope} \end{aligned}$$

$$\therefore \quad \Sigma X_{1\cdot23}^2 = \Sigma X_{1\cdot2} X_{1\cdot23} = \Sigma X_1 X_{1\cdot23}$$

$$\begin{aligned} \text{Again } \Sigma X_{1\cdot34\ldots n} X_{2\cdot34\ldots n} &= \Sigma[(X_1 - b_{13\cdot4\ldots n}X_3 - b_{14\cdot35\ldots n}X_4 - \ldots - b_{1n\cdot34\ldots(n-1)}X_n) X_{2\cdot34\ldots n}] \\ &= \Sigma X_1 X_{2\cdot34\ldots n} \qquad \text{(Using Prope} \end{aligned}$$

Hence the property 2.

Property 3. *The sum of the product of two residuals is zero if all the subscripts (pri as well as secondary) of the one occur among the secondary subscripts of the other, e.g.,*

$$\Sigma X_{1\cdot2} X_{3\cdot12} = \Sigma (X_1 - b_{12}X_2) X_{3\cdot12} = \Sigma X_1 X_{3\cdot12} - b_{12}\Sigma X_2 X_{3\cdot12} = 0 \quad \text{(Using Prope}$$

$$\begin{aligned} \Sigma X_{2\cdot34\ldots n} X_{1\cdot23\ldots n} &= \Sigma[(X_2 - b_{23\cdot4\ldots n}X_3 - b_{24\cdot35\ldots n}X_4 - \ldots - b_{2n\cdot34\ldots(n-1)}X_n) X_{1\cdot23\ldots n}] \\ &= \Sigma X_2 X_{1\cdot23\ldots n} - b_{23\cdot4\ldots n}\Sigma X_3 X_{1\cdot23\ldots n} - b_{24\cdot35\ldots n}\Sigma X_4 X_{1\cdot23\ldots n} \\ &\quad \ldots - b_{2n\cdot34\ldots(n-1)}\Sigma X_n X_{1\cdot} \\ &= 0 \qquad \text{(Using Prope} \end{aligned}$$

Hence the property 3.

12·6·1. Variance of the Residual.

Let us consider the plane of regression $X_1 = b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}X_n$ on $X_2, X_3, \ldots, X_n$ *viz.*,

Since all the X_i's are measured from their respective means, we have

$$E(X_i) = 0;\, i = 1, 2, \ldots, n \quad \Rightarrow \quad E(X_{1\cdot23\ldots n}) = 0$$

Hence the variance of the residual is given by :

$$\begin{aligned} \sigma^2_{1\cdot23\ldots n} &= \frac{1}{N}\Sigma\left[X_{1\cdot23\ldots n} - E(X_{1\cdot23\ldots n})\right]^2 = \frac{1}{N}\Sigma X^2_{1\cdot23\ldots n} \\ &= \frac{1}{N}\Sigma X_{1\cdot23\ldots n} X_{1\cdot23\ldots n} = \frac{1}{N}\Sigma X_1 X_{1\cdot23\ldots n} \qquad \text{(Using Property} \\ &= \frac{1}{N}\Sigma X_1 [X_1 - b_{12\cdot34\ldots n}X_2 - b_{13\cdot24\ldots n}X_3 - \ldots - b_{1n\cdot23\ldots(n-1)}X_n] \\ &= \sigma_1^2 - b_{12\cdot34\ldots n}r_{12}\sigma_1\sigma_2 - b_{13\cdot24\ldots n}r_{13}\sigma_1\sigma_3 - \ldots - b_{1n\cdot23\ldots(n-1)}r_{1n}\sigma_1\sigma_n \end{aligned}$$

$$\Rightarrow \quad \sigma_1^2 - \sigma^2_{1\cdot23\ldots n} = b_{12\cdot34\ldots n}r_{12}\sigma_1\sigma_2 + b_{13\cdot24\ldots n}r_{13}\sigma_1\sigma_3 + \ldots + b_{1n\cdot23\ldots(n-1)}$$

Eliminating the b's in equations (12·22) and (12·16b), we get

$$\begin{vmatrix} \sigma_1^2 - \sigma^2_{1\cdot23\ldots n} & r_{12}\sigma_1\sigma_2 & \ldots & r_{1n}\sigma_1\sigma_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & r_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & \ldots & \sigma_n^2 \end{vmatrix} = 0$$

Dividing $R_1, R_2, \ldots, R_n$, by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively and also $C_1, C_2, \ldots, C_n$ by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively, we get

$$\begin{vmatrix} 1-\frac{\sigma^2_{1.23\ldots n}}{\sigma_1^2} & r_{12} & \cdots & r_{1n} \\ r_{12} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r_{1n} & r_{2n} & \cdots & 1 \end{vmatrix} = 0$$

$$\Rightarrow \quad \begin{vmatrix} 1-\frac{\sigma^2_{1.23\ldots n}}{\sigma_1^2} & r_{12} & \cdots & r_{1n} \\ r_{12}+0 & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r_{1n}+0 & r_{2n} & \cdots & 1 \end{vmatrix} = 0$$

$$\begin{vmatrix} 1 & r_1 & \cdots & r_{1n} \\ r_{12} & 1 & \cdots & r_{2n} \\ \vdots & & & \vdots \\ r_{1n} & r_{2n} & \cdots & 1 \end{vmatrix} - \begin{vmatrix} \frac{\sigma^2_{1.23\ldots n}}{\sigma_1^2} & r_{12} & \cdots & r_{1n} \\ 0 & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & r_{2n} & \cdots & 1 \end{vmatrix} = 0$$

$$\Rightarrow \quad \omega - \frac{\sigma^2_{1.23\ldots n}}{\sigma_1^2}\omega_{11} = 0 \quad \Rightarrow \quad \sigma^2_{1.23\ldots n} = \sigma_1^2 \frac{\omega}{\omega_{11}} \qquad \ldots (12{\cdot}23)$$

Remark. In a tri-variate distribution, $\sigma_{1.23}^2 = \sigma_1^2 \frac{\omega}{\omega_{11}}$ ... (12·23a)

where ω and ω_{11} are defined in (12·12).

12·7. COEFFICIENT OF MULTIPLE CORRELATION

In a tri-variate distribution in which each of the variables X_1, X_2, and X_3 has N observations, the multiple correlation coefficient of X_1 on X_2 and X_3, usually denoted by $R_{1.23}$, is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3 on X_1. In other words $R_{1.23}$ is the correlation coefficient between X_1 and its estimated value as given by the plane of regression of X_1 on X_2 and X_3 *viz.*,

$$e_{1.23} = b_{12.3}X_2 + b_{13.2}X_3$$

We have $X_{1.23} = X_1 - b_{12.3}X_2 - b_{13.2}X_3 = X_1 - e_{1.23} \quad \Rightarrow \quad e_{1.23} = X_1 - X_{1.23}$

Since X_i's are measured from their respective means, we have

$$E(X_{1.23}) = 0 \text{ and } E(e_{1.23}) = 0 \qquad [\because E(X_i) = 0;\ i = 1, 2, 3] \qquad \ldots (*)$$

By def., $$R_{1.23} = \frac{\text{Cov}\,(X_1, e_{1.23})}{\sqrt{V(X_1)\,V(e_{1.23})}} \qquad \ldots (12{\cdot}24)$$

$$\text{Cov}\,(X_1, e_{1.23}) = E\left[\{X_1 - E(X_1)\}\{e_{1.23} - E(e_{1.23})\}\right] = E(X_1\, e_{1.23}) \qquad [\text{From } (*)]$$

$$= \frac{1}{N}\Sigma X_1\, e_{1.23} = \frac{1}{N}\Sigma X_1\,(X_1 - X_{1.23})$$

$$= \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X_1 X_{1.23} = \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X^2_{1.23}$$

(Using Property 2, § 12·6)

$$= \sigma_1^2 - \sigma_{1.23}^2$$

Also $$V(e_{1.23}) = E(e_{1.23}^2) = \frac{1}{N}\Sigma e_{1.23}^2 = \frac{1}{N}\Sigma(X_1 - X_{1.23})^2$$

$$= \frac{1}{N}\Sigma(X_1^2 + X_{1.23}^2 - 2X_1X_{1.23}) = \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma X_{1.23}^2 - \frac{2}{N}\Sigma X_1 X_{1.23}$$

$$= \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma X_{1.23}^2 - \frac{2}{N}\Sigma X_{1.23}^2 \qquad \text{(Using property 2, § 12·6)}$$

$$= \sigma_1^2 - \sigma_{1.23}^2$$

Substituting in (12·24), we get

$$R_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1.23}^2)}} \quad \Rightarrow \quad R^2_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sigma_1^2} = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

$$\Rightarrow \quad 1 - R^2_{1.23} = \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

Using (12·23a), we get

$$1 - R^2_{1.23} = \frac{\omega}{\omega_{11}} \qquad \ldots (12{\cdot}25)$$

where

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \qquad \text{(on simplification)}$$

and $$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$$

Hence from (12·25), we get

$$R^2_{1.23} = 1 - \frac{\omega}{\omega_{11}} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \qquad \ldots (12{\cdot}25a)$$

This formula expresses the multiple correlation coefficient in terms of the total correlation coefficients between the pairs of variables.

Remark. $R^2_{1.23} \le 1 \quad \Rightarrow \quad 1 - R^2_{1.23} = \frac{\omega}{\omega_{11}} \ge 0 \qquad \ldots (*)$

But $\omega_{11} = 1 - r^2_{23} > 0 \qquad (\because |r_{23}| \le 1)$

$\therefore$ From (*), $\omega \ge 0 \qquad \ldots (**)$

Also $\omega_{22} = (-1)^{2+2}(1 - r_{13}^2) \ge 0$ and $\omega_{33} = (-1)^{3+3}(1 - r_{12}^2) \ge 0$

In general, $\omega_{ii} \ge 0\ ;\ i = 1, 2, 3 \qquad (12{\cdot}25b)$

Generalisation. In case of n-variate distribution, the multiple correlation coefficient of X_1 on $X_2, X_3, \ldots, X_n$, usually denoted by $R_{1.23\ldots n}$, is the correlation coefficient between X_1 and $e_{1.23\ldots n} = X_1 - X_{1.23\ldots n}$

$$\therefore \quad R_{1.23\ldots n} = \frac{\text{Cov}\,(X_1, e_{1.23\ldots n})}{\sqrt{V(X_1)\,V(e_{1.23\ldots n})}}$$

$$\text{Cov}\,(X_1, e_{1.23\ldots n}) = \frac{1}{N}\Sigma X_1 e_{1.23\ldots n} = \frac{1}{N}\Sigma X_1(X_1 - X_{1.23\ldots n})$$

$$= \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X_1 X_{1.23\ldots n}$$

$$= \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X^2_{1.23\ldots n} = \sigma_1^2 - \sigma^2_{1.23\ldots n}$$

$$V(e_{1.23\ldots n}) = \frac{1}{N}\Sigma e^2_{1.23\ldots n} = \frac{1}{N}\Sigma (X_1 - X_{1.23\ldots n})^2$$

$$= \frac{1}{N}\Sigma (X_1^2 + X^2_{1.23\ldots n} - 2X_1X_{1.23\ldots n})$$

$$= \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma X^2_{1.23\ldots n} - 2\frac{1}{N}\Sigma X_1X_{1.23\ldots n}$$

$$= \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma X^2_{1.23\ldots n} - \frac{2}{N}\Sigma X^2_{1.23\ldots n}$$

$$= \sigma_1^2 - \sigma^2_{1.23\ldots n}$$

$$\therefore \quad R_{1.23\ldots n} = \frac{\sigma_1^2 - \sigma^2_{1.23\ldots n}}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma^2_{1.23\ldots n})}} = \left(\frac{\sigma_1^2 - \sigma^2_{1.23\ldots n}}{\sigma_1^2}\right)^{1/2}$$

$$R^2_{1.23\ldots n} = 1 - \frac{\sigma^2_{1.23\ldots n}}{\sigma_1^2} = 1 - \frac{\omega}{\omega_{11}}, \qquad \ldots(12\cdot25c)$$

where ω and ω_{11} are defined in (12·18).

Remarks 1. *It may be pointed out here that multiple correlation coefficient can never be negative, because from (*) and (**), we get* $\text{Cov}(X_1, e_{1.23\ldots n}) = \sigma_1^2 - \sigma^2_{1.23\ldots n} = \text{Var}(e_{1.23\ldots n}) \geq 0$

Since the sign of $R_{1.23\ldots n}$ *depends upon the covariance term* $\text{Cov}(X_1, e_{1.23\ldots n})$, *we conclude that* $R_{1.23\ldots n} \geq 0$.

2. Since $R^2_{1.23\ldots n} \geq 0$, we have $1 - \frac{\omega}{\omega_{11}} \geq 0 \Rightarrow \omega \leq \omega_{11}$ ... (12·25*d*)

3. Also $R^2_{1.23\ldots n} \leq 1 \Rightarrow 1 - R^2_{1.23\ldots n} = \frac{\omega}{\omega_{11}} \geq 0 \Rightarrow \omega \geq 0$ ... (12·25*e*)

From the above results, we get $\omega_{11} \geq \omega \geq 0$ ... (12·25*f*)

In general, we have $\omega_{ij} \geq 0\ ;\ i = 1, 2, \ldots, n$

4. Since ω is symmetric in r_{ij}'s, we have $\omega_{ij} = \omega_{ji}\ ;\ i \neq j = 1, 2, \ldots, n$ ... (12·25*g*)

12·7·1. Properties of Multiple Correlation Coefficient

1. Multiple correlation co-efficient measures the closeness of the association between the observed values and the expected values of a variable obtained from the multiple linear regression of that variable on other variables.

2. Multiple correlation coefficient between observed values and expected values, when the expected values are calculated from a linear relation of the variables determined by the method of least squares, is always greater than that where expected values are calculated from any other linear combination of the variables.

3. Since $R_{1.23}$ is the simple correlation between X_1 and $e_{1.23}$, it must lie between –1 and +1. But as seen in Remark 1 above, $R_{1.23}$ is a non-negative quantity we conclude that $0 \leq R_{1.23} \leq 1$.

4. If $R_{1.23} = 1$, then association is perfect and all the regression residuals are zero, and as such $\sigma^2_{1.23} = 0$. In this case, since $X_1^* = e_{1.23}$, the predicted value of X_1, the multiple linear regression equation of X_1 on X_2 and X_3 may be said to be a perfect prediction formula.

5. If $R_{1.23} = 0$, then all total and partial correlations involving X_1 are zero. [See *Example 12·11 (iv)*). So X_1 is completely uncorrelated with all the other variables in this case and the multiple regression equation fails to throw any light on the value of X_1 when X_2 and X_3 are known.

6. $R_{1.23}$ is not less than any total correlation coefficient, *i.e.*, $R_{1.23} \geq r_{12}, r_{13}, r_{23}$.

12·8. COEFFICIENT OF PARTIAL CORRELATION

Sometimes the correlation between two variables X_1 and X_2 may be partly d the correlation of a third variable, X_3 with both X_1 and X_2. In such a situation may want to know what the correlation between X_1 and X_2 would be if the effect on each of X_1 and X_2 were eliminated. This correlation is called the *partial corre* and the correlation coefficient between X_1 and X_2 after the linear effect of X_3 o of them has been eliminated is called the *partial correlation coefficient*.

The residual $X_{1.3} = X_1 - b_{13}X_3$, may be regarded as that part of the variab which remains after the linear effect of X_3 has been eliminated. Similarly, the res $X_{2.3}$ may be interpreted as the part of the variable X_2 obtained after eliminatin linear effect of X_3. Thus the partial correlation coefficient between X_1 and X_2, us denoted by $r_{12.3}$, is given by :

$$r_{12.3} = \frac{\text{Cov}(X_{1.3}, X_{2.3})}{\sqrt{\text{Var}(X_{1.3})\ \text{Var}(X_{2.3})}} \qquad \ldots (1$$

We have

$$\text{Cov}(X_{1.3}, X_{2.3}) = \frac{1}{N}\Sigma X_{1.3}X_{2.3} = \frac{1}{N}\Sigma X_1X_{2.3} = \frac{1}{N}\Sigma X_1(X_2 - b_{23}X_3)$$

$$= \frac{1}{N}\Sigma X_1X_2 - b_{23}\frac{1}{N}\Sigma X_1X_3 = r_{12}\sigma_1\sigma_2 - r_{23}\frac{\sigma_3}{\sigma_3}\cdot(r_{13}\sigma_1\sigma_3)$$

$$= \sigma_1\sigma_2(r_{12} - r_{13}r_{23})$$

$$V(X_{1.3}) = \frac{1}{N}\Sigma X_{1.3}^2 = \frac{1}{N}\Sigma X_{1.3}X_{1.3}$$

$$= \frac{1}{N}\Sigma X_1X_{1.3} = \frac{1}{N}\Sigma X_1(X_1 - b_{13}X_3)$$

$$= \frac{1}{N}\Sigma X_1^2 - b_{13}\cdot\frac{1}{N}\Sigma X_1X_3 = \sigma_1^2 - r_{13}\frac{\sigma_1}{\sigma_3}r_{13}\sigma_1\sigma_3$$

$$= \sigma_1^2(1 - r_{13}^2)$$

Similarly, we shall get $V(X_{2.3}) = \sigma_2^2(1 - r_{23}^2)$

Substituting in (12·26), we get

$$r_{12.3} = \frac{\sigma_1\sigma_2(r_{12} - r_{13}r_{23})}{\sqrt{\sigma_1^2(1 - r_{13}^2)\,\sigma_2^2(1 - r_{23}^2)}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad \ldots (12$$

Aliter. We have

$$0 = \Sigma X_{2.3}X_{1.23} = \Sigma X_{2.3}(X_1 - b_{12.3}X_2 - b_{13.2}X_3)$$

$$= \Sigma X_1X_{2.3} - b_{12.3}\Sigma X_{2.3}X_2 - b_{13.2}\Sigma X_{2.3}X_3 = \Sigma X_{1.3}X_{2.3} - b_{12.3}\Sigma X_{21}X_2$$

$$\therefore \quad b_{12.3} = \frac{\Sigma X_{1.3}X_{2.3}}{\Sigma X_{2.3}^2}$$

From this it follows that $b_{12.3}$ is coefficient of regression of $X_{1.3}$ on $X_{2.3}$

Similarly, $b_{21.3}$ is the coefficient of regression of $X_{2.3}$ on $X_{1.3}$.

Since correlation coefficient is the geometric mean between regression coeffic we have $r^2_{12.3} = b_{12.3} \times b_{21.3}$

But by def., $b_{12\cdot3} = -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}$ and $b_{21\cdot3} = -\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}}$

$$r^2_{12\cdot3} = \left(-\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}\right)\left(-\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}}\right) = \frac{\omega_{12}^2}{\omega_{11}\omega_{22}} \qquad (\because \omega_{12} = \omega_{21})$$

$$\Rightarrow \quad r_{12\cdot3} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}}, \qquad \ldots (12\cdot26b)$$

the negative sign being taken since the sign of regression coefficients is the same as that of $(-\omega_{12})$.

Substituting the values of ω_{12}, ω_{11} and ω_{22} from (12·12), we get

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Remarks 1. The expressions for $r_{13\cdot2}$ and $r_{23\cdot1}$ can be similarly obtained, to give

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\,r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \quad \text{and} \quad r_{23\cdot1} = \frac{r_{23} - r_{21}r_{13}}{\sqrt{(1-r_{21}^2)(1-r_{13}^2)}} \qquad \ldots (12\cdot26c)$$

2. If $r_{12\cdot3} = 0$, we have then $r_{12} = r_{13}\,r_{23}$. It means that r_{12} will not be zero if X_3 is correlated with both X_1 and X_2. Thus, although X_1 and X_2 may be uncorrelated when effect of X_3 is eliminated, yet X_1 and X_2 may appear to be correlated because they carry the effect of X_3 on them.

3. Partial correlation coefficient helps in deciding whether to include or not an additional independent variable in regression analysis.

4. We know that $\sigma_1^2(1-r_{12}^2)$ and $\sigma_1^2(1-r_{13}^2)$ are the residual variances if X_1 is estimated from X_2 and X_3 individually, while $\sigma_1^2(1-R_{1\cdot23}^2)$ is the residual variance if X_1 is estimated from X_2 and X_3 taken together. So from the above remark and $R_{1\cdot23}^2 \geq r_{12}^2$ and r_{13}^2, it follows that inclusion of an additional variable can only reduce the residual variance. Now inclusion of X_3 when X_2 has already been taken for predicting X_1, is worthwhile only when the resultant reduction in the residual variance is substantial. This will be the case when $r_{13\cdot2}$ is sufficiently large. Thus in this respect partial correlation coefficient has its significance in regression analysis.

12·8·1. Generalisation. In the case of n variables $X_1, X_2, \ldots, X_n$ the partial correlation coefficient $r_{12\cdot34\ldots n}$ between X_1 and X_2 (after the linear effect of $X_3, X_4, \ldots, X_n$ on them has been eliminated), is given by :

$$r^2_{12\cdot34\ldots n} = b_{12\cdot34\ldots n} \times b_{21\cdot34\ldots n} \qquad \ldots (12\cdot26d)$$

But, we have

$$\left.\begin{aligned} b_{12\cdot34\ldots n} &= -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \\ \text{and} \quad b_{21\cdot34\ldots n} &= -\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}} \end{aligned}\right\} \quad [c.f.\ \text{Equation (12·20)}]$$

$$r^2_{12\cdot34\ldots n} = \left(-\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}\right)\left(-\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}}\right) = \frac{\omega_{12}^2}{\omega_{11}\omega_{22}} \qquad \ldots (12\cdot26e)$$

$$\Rightarrow \quad r_{12\cdot34\ldots n} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}},$$

negative sign being taken since the sign of the regression coefficient is the same as that

10·9. MULTIPLE CORRELATION IN TERMS OF TOTAL AND PARTIAL CORRELATIONS

$$1 - R_{1\cdot23}^2 = (1-r_{12}^2)(1-r_{13\cdot2}^2) \qquad \ldots (1\cdot26f)$$

Proof. We have

$$1 - R_{1\cdot23}^2 = 1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2} = \frac{1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2}$$

Also $1 - r_{13\cdot2}^2 = 1 - \frac{(r_{13} - r_{12}\,r_{23})^2}{(1-r_{12}^2)(1-r_{23}^2)} = \frac{1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}}{(1-r_{12}^2)(1-r_{23}^2)}$

Hence the result.

Theorem. *Any standard deviation of order 'p' may be expressed in terms of a standar deviation of order (p − 1) and a partial correlation coefficient of order (p − 1).*

Proof. Let us consider the sum :

$$\begin{aligned} \Sigma X^2_{1\cdot23\ldots n} &= \Sigma X_{1\cdot23\ldots n} X_{1\cdot23\ldots n} \\ &= \Sigma[X_{1\cdot23\ldots(n-1)} X_{1\cdot23\ldots n} \qquad \text{(Using Property 2, § 12·} \\ &= \Sigma\big[X_{1\cdot23\ldots(n-1)}\,(X_1 - b_{12\cdot34\ldots n}X_2 - \ldots - b_{1(n-1)\cdot23\ldots n}X_{n-1} - b_{1n\cdot23\ldots(n-1)}X \\ &= \Sigma X_{1\cdot23\ldots(n-1)}\,X_1 - b_{1n\cdot23\ldots(n-1)}\,\Sigma X_{1\cdot23\ldots(n-1)}X_n \quad \text{(Using Property 1 § 1} \\ &= \Sigma X^2_{1\cdot23\ldots(n-1)} - b_{1n\cdot23\ldots(n-1)}\,\Sigma X_{1\cdot23\ldots(n-1)}X_{n\cdot23\ldots(n-1)} \end{aligned}$$

Dividing both sides by N (total number of observations), we get

$$\sigma^2_{1\cdot23\ldots n} = \sigma^2_{1\cdot23\ldots(n-1)} - b_{1n\cdot23\ldots(n-1)}\,\mathrm{Cov}\,[X_{1\cdot23\ldots(n-1)}, X_{n\cdot23\ldots(n-1)}]$$

The regression coefficient of $X_{n\cdot23\ldots(n-1)}$ on $X_{1\cdot23\ldots(n-1)}$ is given by :

$$b_{n1\cdot23\ldots(n-1)} = \frac{\mathrm{Cov}\,[X_{1\cdot23\ldots(n-1)}, X_{n\cdot23\ldots(n-1)}]}{\sigma^2_{1\cdot23\ldots(n-1)}}$$

$$\therefore \quad \sigma^2_{1\cdot23\ldots n} = \sigma^2_{1\cdot23\ldots(n-1)}\left[1 - b_{1n\cdot23\ldots(n-1)}\,b_{n1\cdot23\ldots(n-1)}\right] \qquad \text{[From (*) a}$$

$$= \sigma^2_{1\cdot23\ldots(n-1)}\left[1 - r^2_{n1\cdot23\ldots(n-1)}\right],$$

a formula which expresses the standard deviation of order $(n-1)$ in terms of deviation of order $(n-2)$ and partial correlation coefficient of order $(n-2)$. $p = (n-1)$, the theorem is established.

Cor. 1. From (12·27), we have $\sigma^2_{1\cdot23\ldots(n-1)} = \sigma^2_{1\cdot23\ldots(n-2)}\left[1 - r^2_{1(n-1)\cdot23\ldots(n-2)}\right]$

and so on. Thus the repeated application of (12·27) gives

$$\sigma^2_{1\cdot23\ldots n} = \sigma_1^2(1-r_{12}^2)\,(1-r_{13\cdot2}^2)\,(1-r_{14\cdot32}^2)\ldots(1-r^2_{1n\cdot23\ldots(n-1)})$$

Since partial correlation coefficients cannot exceed unity numerically, we ge

(12·27*a*) and so on :

$$\left.\begin{aligned} \sigma^2_{1\cdot23\ldots n} &\leq \sigma^2_{1\cdot23\ldots(n-1)} \\ \sigma^2_{1\cdot23\ldots(n-1)} &\leq \sigma^2_{1\cdot23\ldots(n-2)} \\ &\vdots \qquad\quad \vdots \\ \sigma^2_{1\cdot23} &\leq \sigma^2_{1\cdot2} \\ \sigma^2_{1\cdot2} &\leq \sigma_1^2 \end{aligned}\right\}$$

$$\Rightarrow \qquad \sigma_1 \geq \sigma_{1.2} \geq \sigma_{1.23} \geq \quad \ldots \quad \geq \sigma_{1.23\ldots n} \qquad \ldots (12\cdot27c)$$

Cor. 2. Also, we have $\sigma^2_{1.23\ldots n} = \sigma_1^2(1 - R^2_{1.23\ldots n})$

On using (12·27b), we get

$$1 - R^2_{1.23\ldots n} = (1 - r_{12}^2)(1 - r_{13.2}^2)\ldots(1 - r^2_{1n.23\ldots(n-1)}) \qquad \ldots (12\cdot27d)$$

This is the generalisation of the result obtained in (12·26f)

Since $\quad |r_{ij.k}| \leq 1;\ k = 0, 1, 2, \ldots, (n-1),$

where $r_{ij.k}$ is a partial correlation coefficient of order s, we get from (12·27d)

$$1 - R^2_{1.23\ldots n} \leq 1 - r_{12}^2,\ 1 - R^2_{1.23\ldots n} \leq 1 - r^2_{13.2} \quad \text{and so on.}$$

i.e., $\quad R^2_{1.23\ldots n} \geq r_{12}^2,\ r^2_{13.2}, \ldots, r^2_{1n.23\ldots(n-1)}$ $\qquad \ldots (12\cdot27e)$

Since $R_{1.23\ldots n}$ is symmetric in its secondary subscripts, we have

$$\left.\begin{aligned} R^2_{1.23\ldots n} &\geq r_{1i}^2,\ (i = 2, 3, \ldots, n) \\ R^2_{1.23\ldots n} &\geq r_{ij}^2\ (i \neq j = 2, 3, \ldots, n) \end{aligned}\right\} \qquad \ldots (12\cdot27f)$$

and so on.

12·10. EXPRESSION FOR REGRESSION COEFFICIENT IN TERMS OF REGRESSION COEFFICIENTS OF LOWER ORDER

Consider

$$\begin{aligned} \Sigma X_{1.34\ldots n} X_{2.34\ldots n} &= \Sigma X_{1.34\ldots(n-1)} X_{2.34\ldots n} \\ &= \Sigma X_{1.34\ldots(n-1)}(X_2 - b_{23.4\ldots n} X_3 - \ldots - b_{2n.34\ldots(n-1)} X_n) \\ &= \Sigma X_{1.34\ldots(n-1)} X_2 - b_{2n.34\ldots(n-1)} \Sigma X_{1.34\ldots(n-1)} X_n \\ &= \Sigma X_{1.34\ldots(n-1)} X_{2.34\ldots(n-1)} - b_{2n.34\ldots(n-1)} \Sigma X_{1.34\ldots(n-1)} X_{n.34\ldots(n-1)} \end{aligned}$$

Dividing both sides by N, the total number of observations, we get

$$\begin{aligned} Cov\,(X_{1.34\ldots n}, X_{2.34\ldots n}) &= Cov\,(X_{1.34\ldots(n-1)}, X_{2.34\ldots(n-1)}) \\ &\quad - b_{2n.34\ldots(n-1)}\, Cov\,(X_{1.34\ldots(n-1)}, X_{n.34\ldots(n-1)}) \end{aligned}$$

$$\Rightarrow \quad b_{12.34\ldots n}\,\sigma^2_{2.34\ldots n} = b_{12.34\ldots(n-1)}\,\sigma^2_{2.34\ldots(n-1)} - b_{2n.34\ldots(n-1)}\, b_{1n.34\ldots(n-1)}\,\sigma^2_{n.34\ldots(n-1)}$$

On using (12·27), we get

$$b_{12.34\ldots n}\,\sigma^2_{2.34\ldots(n-1)}\left[1 - r^2_{2n.34\ldots(n-1)}\right]$$

$$= \sigma^2_{2.34\ldots(n-1)}\left[b_{12.34\ldots(n-1)} - b_{2n.34\ldots(n-1)}\, b_{1n.34\ldots(n-1)} \times \frac{\sigma^2_{n.34\ldots(n-1)}}{\sigma^2_{2.34\ldots(n-1)}}\right] \quad \ldots (*)$$

In the case of two variables, we have

$$b_{ij}\,\sigma_j^2 = b_{ji}\,\sigma_i^2 = Cov\,(X_i, X_j) \quad \Rightarrow \quad b_{ij} = \frac{\sigma_i^2}{\sigma_j^2}\, b_{ji}$$

$$\therefore \quad b_{2n.34\ldots(n-1)}\,\frac{\sigma^2_{n.34\ldots(n-1)}}{\sigma^2_{2.34\ldots(n-1)}} = b_{n2.34\ldots(n-1)}$$

Hence from (*), we get

$$b_{12.34\ldots n}\left[1 - r^2_{2n.34\ldots(n-1)}\right] = \left[b_{12.34\ldots(n-1)} - b_{1n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}\right]$$

$$\Rightarrow \quad b_{12.34\ldots n} = \left[\frac{b_{12.34\ldots(n-1)} - b_{1n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}}{1 - r^2_{2n.34\ldots(n-1)}}\right] \qquad \ldots (12\cdot3)$$

$$\therefore \quad b_{12.34\ldots n} = \left[\frac{b_{12.34\ldots(n-1)} - b_{1n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}}{1 - b_{2n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}}\right] \qquad \ldots (12\cdot3a)$$

Hence, the result.

12·11. EXPRESSION FOR PARTIAL CORRELATION COEFFICIENT IN TERMS OF PARTIAL CORRELATION COEFFICIENT OF LOWER ORDER

By def., $\quad b_{ij.34\ldots n} = r_{ij.34\ldots n} \times \dfrac{\sigma_{i.34\ldots n}}{\sigma_{j.34\ldots n}} \qquad \ldots (12\cdot3)$

$$\begin{aligned} \therefore \quad b_{1n.34\ldots(n-1)} b_{n2.34\ldots(n-1)} &= r_{1n.34\ldots(n-1)} \frac{\sigma_{1.34\ldots(n-1)}}{\sigma_{n.34\ldots(n-1)}} \times r_{n2.34\ldots(n-1)} \frac{\sigma_{n.34\ldots(n-1)}}{\sigma_{2.34\ldots(n-1)}} \\ &= r_{1n.34\ldots(n-1)}\, r_{n2.34\ldots(n-1)} \frac{\sigma_{1.34\ldots(n-1)}}{\sigma_{2.34\ldots(n-1)}} \end{aligned} \qquad \ldots (12\cdot3)$$

Hence from (12·28), on using (12·29) and (12·29a), we get

$$\begin{aligned} r_{12.34\ldots n} \times \frac{\sigma_{1.34\ldots n}}{\sigma_{2.34\ldots n}} &= \left[\frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\, r_{n2.34\ldots(n-1)}}{1 - r^2_{2n.34\ldots(n-1)}}\right] \frac{\sigma_{1.34\ldots(n-1)}}{\sigma_{2.34\ldots(n-1)}} \end{aligned} \qquad \ldots (12\cdot3)$$

Also on using (12·27), we get

$$\frac{\sigma_{1.34\ldots n}}{\sigma_{2.34\ldots n}} = \frac{\sigma_{1.34\ldots(n-1)}}{\sigma_{2.34\ldots(n-1)}} \times \left[\frac{1 - r^2_{1n.34\ldots(n-1)}}{1 - r^2_{2n.34\ldots(n-1)}}\right]^{1/2}$$

Hence from (12·29b), we get

$$r_{12.34\ldots n}\left[\frac{1 - r^2_{1n.34\ldots(n-1)}}{1 - r^2_{2n.34\ldots(n-1)}}\right]^{\frac{1}{2}} = \left[\frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\, r_{n2.34\ldots(n-1)}}{1 - r^2_{2n.34\ldots(n-1)}}\right]$$

$$\Rightarrow \quad r_{12.34\ldots n} = \frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\, r_{n2.34\ldots(n-1)}}{[1 - r^2_{1n.34\ldots(n-1)}]^{1/2}\,[1 - r^2_{n2.34\ldots(n-1)}]^{1/2}}$$

which is an expression for the correlation coefficient of order $p = (n-2)$ in terms of correlation coefficient of order $(p-1) = (n-3)$.

Example 12·7 *From the data relating to the yield of dry bark (X_1), height (X_2) and girth (X_3) for 18 cinchona plants, the following correlation coefficients were obtained*

$$r_{12} = 0{\cdot}77,\ r_{13} = 0{\cdot}72 \quad \text{and} \quad r_{23} = 0{\cdot}52$$

Find the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient

Solution. $\quad r_{12.3} = \dfrac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \dfrac{0{\cdot}77 - 0{\cdot}72 \times 0{\cdot}52}{\sqrt{[1 - (0{\cdot}72)^2][1 - (0{\cdot}52)^2]}} = 0$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}$$

$$= \frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - (0.52)^2} = 0.7334$$

$R_{1.23} = +0.8564$ (since multiple correlation coefficient is non-negative).

Example 12·8. *In a trivariate distribution :*

$\sigma_1 = 2,\ \sigma_2 = \sigma_3 = 3,\ r_{12} = 0.7,\ r_{23} = r_{31} = 0.5.$

Find (i) $r_{23.1}$, (ii) $R_{1.23}$, (iii) $b_{12.3}$, $b_{13.2}$, and (iv) $\sigma_{1.23}$.

Solution. We have

(i) $$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{0.5 - (0.7)(0.5)}{\sqrt{(1 - 0.49)(1 - 0.25)}} = 0.2425$$

(ii) $$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.49 + 0.25 - 2(0.7)(0.5)(0.5)}{1 - 0.25} = 0.52 \qquad \therefore \quad R_{1.23} = +0.7211.$$

(iii) $$b_{12.3} = r_{12.3}\frac{\sigma_{1.3}}{\sigma_{2.3}} \quad \text{and} \quad b_{13.2} = r_{13.2}\frac{\sigma_{1.2}}{\sigma_{3.2}} \qquad \ldots (*)$$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = 0.6, \qquad r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = 0.2425$$

$$\sigma_{1.3} = \sigma_1\sqrt{(1 - r_{13}^2)} = 2\sqrt{(1 - 0.25)} = 1.7320$$

$$\sigma_{2.3} = \sigma_2\sqrt{(1 - r_{23}^2)} = 3\sqrt{(1 - 0.25)} = 2.5980$$

$$\sigma_{1.2} = \sigma_1\sqrt{(1 - r_{12}^2)} = 2\sqrt{(1 - 0.49)} = 1.4282$$

$$\sigma_{3.2} = \sigma_3\sqrt{(1 - r_{32}^2)} = 3\sqrt{(1 - 0.25)} = 2.5980$$

Substituting in (*), we get :

$$b_{12.3} = 0.6 \times \frac{1.7320}{2.5980} = 0.4 \quad \text{and} \quad b_{13.2} = 0.2425 \times \frac{1.4282}{2.5980} = 0.1333$$

(iv) $$\sigma_{1.23} = \sigma_1\sqrt{\frac{\omega}{\omega_{11}}}$$

where $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} = 0.36$$

and $$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - 0.25 = 0.75$$

$$\therefore \quad \sigma_{1.23} = 2 \times \sqrt{\frac{0.36}{0.75}} = 2 \times \sqrt{0.48} = 2 \times 0.6928 = 1.3856.$$

Example 12·9. *Find the regression equation of X_1 on X_2 and X_3 given the following results :*

Trait	*Mean*	*Standard deviation*	r_{12}	r_{23}	r_{31}
X_1	28·02	4·42	+0·80	—	—
X_2	4·91	1·10	—	−0·56	—
X_3	594	85	—	—	−0·40

where X_1 = *Seed per acre;* X_2 = *Rainfall in inches*

X_3 = *Accumulated temperature above 42°F*

Solution. Regression equation of X_1 on X_2 and X_3 is given by :

$$(X_1 - \bar{X}_1)\frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2)\frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3)\frac{\omega_{13}}{\sigma_3} = 0$$

where $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0.56)^2 = 0.686$$

$$\omega_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}r_{23} - r_{21} = -0.576$$

$$\omega_{13} = r_{23}r_{12} - r_{13} = (-0.56)(0.80) - (-0.40) = -0.048$$

$\therefore$ Required equation of plane of regression of X_1 on X_2 and X_3 is given by :

$$\frac{0.686}{4.42}(X_1 - 28.02) + \frac{(-0.576)}{1.10}(X_2 - 4.91) + \frac{(-0.048)}{85.00}(X_3 - 594) = 0.$$

Example 12·10. *Five hundred students were examined in three subjects I, II and III, each subject carrying 100 marks. A student getting 120 or more but less than 150 marks was put in pass class. A student getting 150 or more but less than 180 marks was put in second class and a student getting 180 or more marks was put in the first class. The following marks were obtained :*

	I	II	II
Mean :	35·8	52·4	48·8
S.D. :	4·2	5·3	6·1
Correlation :	$r_{12} = 0.6$,	$r_{13} = 0.7$	$r_{23} = 0.8$

Assuming that the marks are normally distributed,

(i) *Find the number of students in each of the three classes.*

(ii) *Find the total number of students with total marks lying between 120 and 190.*

(iii) *Find the probability that a student gets more that 240 marks.*

(iv) *What should be the correlation between marks in subjects I and II among students who scored equal marks in subject III ?*

(v) *If r_{23} was not known, obtain the limits within which it may lie from the values of r_{12} and r_{13} (ignoring sampling errors).*

Solution. If Z denotes the total marks of the students in the three subjects and X_1, X_2, X_3 the total marks of the students in subjects I, II and III respectively, then

$$Z = X_1 + X_2 + X_3 \quad \Rightarrow \quad E(Z) = E(X_1) + E(X_2) + E(X_3) = 35.8 + 52.4 + 48.8 =$$

137

$$V(Z) = V(X_1) + V(X_2) + V(X_3) + 2[\text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_3) + \text{Cov}(X_3, X_1)]$$

$$= 17.64 + 28.09 + 37.21 + 26.712 + 35.868 + 51.728$$

$$= 197.248 \qquad [\text{Using Cov}(X_i, X_j) = r_{ij}\sigma_i\sigma_j]$$

$\Rightarrow \quad \sigma_Z = \sqrt{197{\cdot}248} = 14{\cdot}045$

Now $\xi = \dfrac{Z - E(Z)}{\sigma_Z} \sim N(0, 1)$

Z	$\xi = \frac{Z-137}{14{\cdot}045}$	$p = \int_{-\infty}^{\xi} p(\xi)d\xi$	Class	Area under the curve in this class (A)	Frequency 500 × (A)
120 –	1·21050	0·11314	120 – 150	0·70937	354·685
150	0·92567	0·82251	150 – 180	0·17639	88·195
180	3·06180	0·99890	180 –	0·00102	0·510
190	3·77400	0·99992	120 – 190	0·88678	443·390
240	7·33410	1·00000	240 –	0·00000	0·000

(*i*) The number of students in the pass, second and first class respectively are 355, 88 and 0 (approx.)

(*ii*) Total number of students with total marks between 120 and 190 is 443.

(*iii*) Probability that a student gets more than 240 marks is zero.

(*iv*) The correlation coefficient between marks in subjects I and II of the students who secured equal marks in subject III is $r_{12\cdot3}$ and is given by

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0{\cdot}04}{\sqrt{(1 - 0{\cdot}49)(1 - 0{\cdot}64)}} = 0{\cdot}0934.$$

(*v*) We have

$$r_{12\cdot3}^2 = \frac{(r_{12} - r_{13}\,r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)} \le 1$$

$$\therefore \quad \frac{(0{\cdot}6 - 0{\cdot}7a)^2}{(1 - 0{\cdot}49)(1 - a^2)} \le 1, \text{ where } a = r_{23}$$

$$\Rightarrow \quad 0{\cdot}36 + 0{\cdot}49a^2 - 0{\cdot}84a \le 0{\cdot}51\,(1 - a^2)$$

$$\Rightarrow \quad a^2 - 0{\cdot}84a - 0{\cdot}15 \le 0$$

Thus 'a' lies between the roots of the equation : $a^2 - 0{\cdot}84a - 0{\cdot}15 = 0$, which are 0·99 and −0·15.

Hence r_{23} should lie between −0·15 and 0·99.

Example 12·11. *Show that* $1 - R_{1\cdot23}^2 = (1 - r_{12}^2)\,(1 - r_{13\cdot2}^2)$ ... (*)

Deduce that (*i*) $R_{1\cdot23} \ge r_{12}$, (*ii*) $R_{1\cdot23}^2 = r_{12}^2 + r_{13}^2$, *if* $r_{23} = 0$.

(*iii*) $1 - R_{1\cdot23}^2 = \dfrac{(1-\rho)(1+2\rho)}{(1+\rho)}$, *provided all coefficients of zero order are equal to* ρ.

(*iv*) *If* $R_{1\cdot23} = 0$, X_1 *is uncorrelated with any of other variables, i.e.,* $r_{12} = r_{13} = 0$.

Solution. For the proof of (*), see (12·26*f*), § 12·9.

(*i*) Since $|\,r_{13\cdot2}\,| \le 1$, we have from (*)

$$1 - R_{1\cdot23}^2 \le 1 - r_{12}^2 \quad \Rightarrow \quad R_{1\cdot23} \ge r_{12}$$

(*ii*) We have $r_{13\cdot2} = \dfrac{r_{13} - r_{12}\,r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \dfrac{r_{13}}{\sqrt{1 - r_{12}^2}}$, (if $r_{23} = 0$)

$\therefore$ From (*), we get

$$1 - R_{1\cdot23}^2 = (1 - r_{12}^2)\left[\,1 - \frac{r_{13}^2}{1 - r_{12}^2}\right] = 1 - r_{12}^2 - r_{13}^2$$

Hence $R_{1\cdot23}^2 = r_{12}^2 + r_{13}^2$, if $r_{23} = 0$.

(*iii*) Here, we are given that $r_{12} = r_{13} = r_{23} = \rho$

$$\therefore \quad r_{13\cdot2} = \frac{\rho - \rho^2}{\sqrt{(1-\rho^2)(1-\rho^2)}} = \frac{\rho(1-\rho)}{(1-\rho^2)} = \frac{\rho}{1+\rho}$$

Hence from (*), we have

$$1 - R_{1\cdot23}^2 = (1 - \rho^2)\left[1 - \frac{\rho^2}{(1+\rho)^2}\right] = \frac{(1-\rho)(1+2\rho)}{(1+\rho)}$$

(*iv*) If $R_{1\cdot23} = 0$, (*) gives : $1 = (1 - r_{12}^2)(1 - r_{13\cdot2}^2)$

Since $0 \le r_{12}^2 \le 1$ and $0 \le r_{13\cdot2}^2 \le 1$, (**) will hold if and only if $r_{12} = 0$ and $r_{13\cdot2} = 0$

Now $r_{13\cdot2} = 0 \;\Rightarrow\; \dfrac{r_{13} - r_{12}\,r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = 0 \;\Rightarrow\; \dfrac{r_{13}}{\sqrt{1 - r_{32}^2}} = 0$

$\therefore \quad r_{13} = 0$

Thus if $R_{1\cdot23} = 0$, then $r_{13} = r_{12} = 0$, *i.e.*, X_1 is uncorrelated with X_2 and X_3.

Example 12·12. *Show that the correlation coefficient between the residuals* $X_{1\cdot23}$ *and* $X_{2\cdot13}$ *is equal and opposite to that between* $X_{1\cdot3}$ *and* $X_{2\cdot3}$.

Solution. The correlation coefficient between $X_{1\cdot23}$ and $X_{2\cdot13}$ is given by :

$$\frac{\mathrm{Cov}\,(X_{1\cdot23}, X_{2\cdot13})}{\sigma_{1\cdot23}\,\sigma_{2\cdot13}} = \frac{\sum X_{1\cdot23}\,X_{2\cdot13}}{N\,\sigma_{1\cdot23}\,\sigma_{2\cdot13}} = \frac{\frac{1}{N}\sum X_{2\cdot13}(X_1 - b_{12\cdot3}\,X_2 - b_{13\cdot2}\,X_3)}{\sigma_{1\cdot23}\,\sigma_{2\cdot13}}$$

$$= -\,b_{12\cdot3}\,\frac{\sum X_{2\cdot13}\,X_2}{N\,\sigma_{1\cdot23}\,\sigma_{2\cdot13}} \qquad \text{(c.f. Property 1)}$$

$$= -\,b_{12\cdot3}\,\frac{\sum X_{2\cdot13}^2}{N\sigma_{1\cdot23}\,\sigma_{2\cdot13}} \qquad \text{(c.f. Property 2)}$$

$$= -\,b_{12\cdot3}\,\frac{\sigma_{2\cdot13}}{\sigma_{1\cdot23}} = -\,b_{12\cdot3}\,\frac{(\sigma_2\sqrt{\omega/\omega_{22}})}{(\sigma_1\sqrt{(\omega/\omega_{11})})},$$

where $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 \quad \text{and} \quad \omega_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix}$$

$$\therefore \quad r(X_{1\cdot23}, X_{2\cdot13}) = -\,b_{12\cdot3}\,\frac{\sigma_2}{\sigma_1}\cdot\sqrt{\frac{1 - r_{23}^2}{1 - r_{13}^2}} = -\,b_{12\cdot3}\,\frac{\sigma_{2\cdot3}}{\sigma_{1\cdot3}}$$

$$\left[\text{since } \sigma_{2\cdot3}^2 = \sigma_2^2\,(1 - r_{23}^2) \text{ and } \sigma_{1\cdot3}^2 = \sigma_1^2\,(1 - r_{13}^2)\right]$$

$$\therefore \quad r(X_{1\cdot23}, X_{2\cdot13}) = -\,\frac{\mathrm{Cov}\,(X_{1\cdot3}, X_{2\cdot3})}{\sigma_{2\cdot3}^2}\cdot\frac{\sigma_{2\cdot3}}{\sigma_{1\cdot3}} = -\,\frac{\mathrm{Cov}\,(X_{1\cdot3}, X_{2\cdot3})}{\sigma_{2\cdot3}\,\sigma_{1\cdot3}} = -\,r(X_{1\cdot3}, X_{2\cdot3})$$

Hence the result.

Example 12·13. *Show that if* $X_3 = aX_1 + bX_2$, *the three partial correlations are numerically equal to unity,* $r_{13\cdot2}$ *having the sign of* a, $r_{23\cdot1}$ *the sign of* b *and* $r_{12\cdot3}$ *the sign of* a/b.

$$\omega = [1 + (p-1)\rho]\,(1-\rho)^{p-1}$$

Similarly, we will have

$$\omega_{11} = [1 + (p-2)\rho]\,(1-\rho)^{p-2}$$

$$1 - R^2 = \frac{\omega}{\omega_{11}} = (1-\rho)\left[\frac{1+(p-1)\rho}{1+(p-2)\rho}\right]$$

Example 12·14. *In a p-variate distribution, all the total (order zero) correlation coefficients are equal to $\rho_0 \neq 0$. Let ρ_k denote the partial correlation coefficient of order k and R_k the multiple correlation coefficient of one variate on k other variates. Prove that*

(i) $\rho_0 \geq -\frac{1}{(p-1)}$, *(ii)* $\rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}$, *and* *(iii)* $R_k^2 = \frac{k\rho_0^2}{1+(k-1)\rho_0}$.

Solution. *(i)* We have proved in *Example 12·13*, that $\rho_k = \frac{\rho_0}{1+k\rho_0}$.

In the case of p-variate distribution, the partial correlation coefficient of the highest order is ρ_{p-2} and is given by :

$$\rho_{p-2} = \frac{\rho_0}{1+(p-2)\rho_0}$$

Since $|\rho_{p-2}| \leq 1 \quad \Rightarrow \quad -1 \leq \rho_{p-2} \leq 1$, we have (on considering the lower limit)

$$-1 \leq \frac{\rho_0}{1+(p-2)\rho_0} \quad \Rightarrow \quad -[1+(p-2)\rho_0] \leq \rho_0 \quad \Rightarrow \quad \rho_0 \geq -\frac{1}{(p-1)}$$

(ii) L.H.S. $= \rho_k - \rho_{k-1} = \frac{\rho_0}{1+k\rho_0} - \frac{\rho_0}{1+(k-1)\rho_0}$

$$= \rho_0\left[\frac{-\rho_0}{(1+k\rho_0)(1+(k-1)\rho_0}\right]$$

$$= -\left(\frac{\rho_0}{1+k\rho_0}\right)\left(\frac{\rho_0}{1+(k-1)\rho_0}\right) = -\rho_k\rho_{k-1}$$

(iii) Taking $\rho = \rho_0$ and $k = p-1$ in *part (ii) Example 12·13*, we get

$$1 - R_k^2 = (1-\rho_0)\left[\frac{1+k\rho_0}{1+(k-1)\rho_0}\right]$$

$$\therefore \quad R_k^2 = 1 - \frac{(1-\rho_0)(1+k\rho_0)}{1+(k-1)\rho_0} = \frac{k\rho_0^2}{1+(k-1)\rho_0} \qquad \text{(On simplification)}$$

Example 12·15. *If r_{12} and r_{13} are given, show that r_{23} must lie in the range :*

$$r_{12}\,r_{13} \pm (1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)^{1/2}$$

If $r_{12} = k$ and $r_{13} = -k$, show that r_{23} will lie between -1 and $1 - 2k^2$.

Solution. We have $r_{12\cdot3}^2 = \left[\frac{r_{12} - r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)\,(1-r_{23}^2)}}\right] \leq 1$

$$\therefore \quad (r_{12} - r_{13}r_{23})^2 \leq (1-r_{13}^2)(1-r_{23}^2)$$

$$\Rightarrow \quad r_{12}^2 + r_{13}^2 r_{23}^2 - 2r_{12}\,r_{13}\,r_{23} \leq 1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2 \qquad \ldots (*)$$

$$\Rightarrow \quad r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}\,r_{13}\,r_{23} \leq 1$$

Thus condition holds for consistent values of r_{12}, r_{13} and r_{23}. (*) may be rewritten as :

$$r_{23}^2 - (2r_{12}r_{13})r_{23} + (r_{12}^2 + r_{13}^2 - 1) \leq 0.$$

Hence, for given values of r_{12} and r_{13}, r_{23} must lie between the roots of the quadratic (in r_{23}) equation :

$$r_{23}^2 - (2r_{12}\,r_{13})\,r_{23} + (r_{12}^2 + r_{13}^2 - 1) = 0,$$

which are given by :

$$r_{23} = r_{12}r_{13} \pm \sqrt{r_{12}^2 r_{13}^2 - (r_{12}^2 + r_{13}^2 - 1)}$$

Hence

$$r_{12}\,r_{13} - \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} \leq r_{23} \leq r_{12}r_{13} + \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} \qquad \ldots (**)$$

In other words, r_{23} must lie in the range $r_{12}\,r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2}$

In particular, if $r_{12} = k$ and $r_{13} = -k$, we get from (**)

$$-k^2 - \sqrt{1-k^2-k^2+k^4} \leq r_{23} \leq -k^2 + \sqrt{1-k^2-k^2+k^4}$$

$$\Rightarrow \quad -k^2 - (1-k^2) \leq r_{23} \leq -k^2 + (1-k^2)$$

$$\therefore \quad -1 \leq r_{23} \leq 1 - 2k^2.$$

CHAPTER CONCEPTS QUIZ

1. State which of the following statements is TRUE or FALSE.
   - *(i)* Correlation ratio is used to measure the extent of relationship between non-linearly related variables.
   - *(ii)* Correlation ratio lies between -1 and 1.
   - *(iii)* Correlation ratio is independent of classification of data.
   - *(iv)* The absolute value of the correlation ratio can never be less than the absolute value of the correlation coefficient.
   - *(v)* If $R_{1\cdot23} = 1$, the multiple linear regression of X_1 on X_2 and X_3 is considered as perfect for predictions.
   - *(vi)* If $R_{1\cdot23} = 0$, all regression residuals are zero.
2. Indicate the correct answer :
   - *(i)* Let η_{YX} be the correlation ratio of Y on X and let η_{XY} be the correlation ratio of X on Y. Let ρ_{XY} be the correlation coefficient of X on Y and ρ_{YX} be the correlation coefficient of Y on X, then :
     - *(a)* ρ_{XY} and η_{XY}, and ρ_{YX} & η_{YX} go together.
     - *(b)* ρ_{XY} and ρ_{YX} are same but η_{YX} and η_{XY} are generally different.
     - *(c)* ρ_{XY}, ρ_{YX}, η_{YX} and η_{XY} are the same. *(d)* ρ_{XY}, ρ_{YX}, η_{YX} and η_{XY} are different.
   - *(ii)* Let η_{YX} be the correlation ratio of Y on X then while the absolute value of the correlation ratio can never be less than the absolute value of the correlation coefficient, the η^2_{YX}
     - *(a)* like correlation coefficient, is independent of origin and scale.
     - *(b)* like regression coefficient, is independent of origin and scale.
     - *(c)* is independent of scale measurements and not of origin.
     - *(d)* is neither independent of origin nor scale.
   - *(iii)* Let ρ be the correlation coefficient between X and Y, η_{YX} be the correlation ratio of Y on X and η_{XY} be the correlation ratio of X on Y. If the plot of X and Y lies on a curve such that no ordinate cuts it more than once and the plot is symmetrical about Y axis, then :
     - *(a)* $\rho = 0, \eta_{XY} = 0, \eta_{YX} = 0$. *(b)* $\rho = 0, \eta_{XY} = 0, \eta_{YX} = 1$.
     - *(c)* $\rho = 0, \eta_{XY} = 1, \eta_{YX} = 0$. *(d)* $\rho = 0, \eta_{XY} = 1, \eta_{YX} = 1$.
   - *(iv)* If plot of X and Y is a curve such that no ordinate cuts it more than once, then, for η_{YX} to be correlation ratio of Y on X and η_{XY} to be the correlation ratio of X on Y:
     - *(a)* $\eta_{YX} = 1$ *(b)* $\eta_{YX} = -1$ *(c)* $\eta_{XY} = 1$ *(d)* $\eta_{YX} = 0$.

12-43. (*a*) If X_1, X_2 and X_3 are three variables measured from their respective means as origin and if e_1 is the expected value of X_1 for given values of X_2 and X_3 from the linear regression of X_1 on X_2 and X_3, prove that $\text{Cov}(X_1, e_1) = \text{Var}(e_1) = \text{Var}(X_1) - \text{Var}(X_1 - e_1)$

(*b*) If $r_{12} = k$ and $r_{23} = -k$, show that r_{13} will lie between -1 and $1 - 2k^2$.

12-44. For three variables X, Y and Z, prove that $r_{XY} + r_{YZ} + r_{ZX} \geq -\frac{3}{2}$.

12-45. If the regression of X_1 on $X_2, \ldots, X_p$ is given by :

$$E(X_1 / X_2, \ldots, X_p) = \alpha + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

and $$\begin{vmatrix} \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{32} & \sigma_{33} & \cdots & \sigma_{3p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_{pp} \end{vmatrix} > 0, \begin{pmatrix} \sigma_{ii} = \text{variances} \\ \sigma_{ij} = \text{covariances} \end{pmatrix}$$

then the constants $\alpha, \beta_2, \ldots, \beta_p$ are given by :

$$\alpha = \mu_1 + \frac{R_{12}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_2} \cdot \mu_2 + \frac{R_{13}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_3} \mu_3 + \ldots + \frac{R_{1p}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_p} \cdot \mu_p \text{ and } \beta_j = -\frac{R_{1j}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_j}, \ (j = 1, 2, \ldots, p)$$

where R_{ij} is the cofactor of ρ_{ij} in the determinant (R) of the correlation matrix

$$R = \begin{vmatrix} \rho_{11} & \rho_{12} \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_p{}^2 & \rho_{pp} \end{vmatrix}$$

12-46. Let the maximum correlation between X_0 and any linear function of $X_1, X_2, \ldots, X_n$ be R and if $r_{01} = r_{02} = \ldots = r_{0n} = r$ and all other correlation coefficients are equal to s then show that :

$$R = r\left[\frac{n}{1 + (n-1)s}\right]^{1/2}$$

12-47. X_1, X_2, X_3 have joint distribution as a multinomial distribution with parameters N, p_1, p_2, p_3. If r_{ij} is the correlation coefficient between X_i and X_j, find the expression for r_{12}, r_{23} and r_{31} and hence deduce the expression for the partial correlation coefficient $r_{1.23}$.

12-48. (*i*) If all the inter-correlations between $(p+1)$variates $X_0, X_1, X_2, \ldots, X_p$ are equal to r, show that each of the partial correlation co-efficients of order $p-1$ is equal to $r/[1 + (p-1)r]$ and that the multiple correlation of X_0 on $X_1, X_2, \ldots, X_p$ is given by :

$$1 - R^2_{0(12\ldots p)} = \frac{(1-r)(1-pr)}{1 + (p-1)r}$$

(*ii*) $r_{12} = (r_{12.3} - r_{13.2}\, r_{23.1})/[(1 - r^2_{13.2})^{1/2} (1 - r^2_{23.1})^{1/2}]$.

12-49. If R denotes the multiple correlation co-efficient of X_1 on $X_2, X_3, \ldots, X_p$ in p-variate distribution, prove that

(*i*) $R^2 \geq R_0^2$, where R_0 the correlation of X_1 with any arbitrary linear function of $X_2, X_3, \ldots, X_p$.

(*ii*) $R^2 \geq R_1^2$, where R_1 is the multiple correlation coefficient of X_1 with $X_2, X_3, \ldots, X_k$, $k < p$

(*iii*) $1 - R^2 = \prod_{j=2}^{p} [1 - r^2_{1j.23\ldots(j-1)}]$

12-50. Let X_1 and X_2 be random variables with means 0 and variances 1 and correlation coefficient ρ. Show that :

$$E[\max(X_1^2, X_2^2)] \leq 1 + \sqrt{1 - \rho^2}$$

Using the above inequality, show that for random variables X_1 and X_2 with means μ_1 and variances σ_1^2 and σ_2^2 and correlation coefficient ρ and for any $k > 0$,

$$P[|X_1 - \mu_1| \geq k\sigma_1 \text{ or } |X_2 - \mu_2| \geq k\sigma_2] \leq \frac{1}{k^2}[1 + \sqrt{1 - \rho^2}]$$

unit - III

X X X Finished

14·1. INTRODUCTION

Before giving the notion of sampling we will first define *population*. In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called *population or universe*. Thus in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

It is obvious that for any statistical investigation complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita (monthly) income of the people in India, we will have to enumerate all the earning individuals in the country, which is rather a very difficult task.

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection (*e.g.*, inspection of crackers, explosive materials, etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, *viz.*, administrative and financial implications, time factor, etc., and we take the help of *sampling*.

A finite subset of statistical individuals in a population is called a sample and the number of individuals in a sample is called the sample size.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilised to approximately determine or estimate the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known as *sampling error* and is inherent and unavoidable in any and every sampling scheme. But sampling results in considerable gains, especially in time and cost, not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day-to day practical life. For example, in a hop we assess the quality of sugar, wheat or any other commodity by taking a andful of it from the bag and then decide to purchase it or not. A housewife ormally tests the cooked products to find if they are properly cooked and contain the oper quantity of salt.

2. TYPES OF SAMPLING

Some of the commonly known and frequently used types of sampling are :

(*i*) *Purposive sampling*, (*ii*) *Random sampling*,
(*iii*) *Stratified sampling*, (*iv*) *Systematic sampling*.

low we will precisely explain these terms, without entering into detailed ion.

2·1. Purposive Sampling. Purposive sampling is one in which the sample selected with definite purpose in view. For example, if we want to give the at the standard of living has increased in the city of New Delhi, we may take ls in the sample from rich and posh localities like Defence Colony, South Golf Links, Jor Bagh, Chanakyapuri, Greater Kailash etc. and ignore the

localities where low income group and the middle class families live. This suffers from the drawback of favouritism and nepotism and does not representative sample of the population.

14·2·2. Random Sampling. In this case the sample units are selected and the drawback of purposive sampling, *viz.*, favouritism or subjective completely overcome. *A random sample is one in which each unit of popula equal chance of being included in it.*

Suppose we take a sample of size n from a finite population of size N. are ${}^{N}C_{n}$ possible samples. A sampling technique in which each of the ${}^{N}C_{n}$ an equal chance of being selected is known as *random sampling* and obtained by this technique is termed as a *random sample*.

Proper care has to be taken to ensure that the selected sample is random bias, which varies from individual to individual, is inherent in any sampli administered by human being. Fairly good random samples can be obtain use of *Tippet's random number tables* or by throwing of a dice, draw of a lotter

The simplest method, which is normally used is the *lottery syste* illustrated below by means of an example.

Suppose we want to select 'r' candidates out of n. We assign number one number to each candidate and write these numbers (1 to n) on n slips made as homogeneous as possible in shape, size, etc. These slips are then pu and thoroughly shuffled and then 'r' slips are drawn one by one. The 'r' corresponding to the numbers on the slips drawn, will constitute the random

Remark. *Tippet's Random Numbers.* L.H.C. Tippet's random number tables consi four-digited numbers, giving in all 10400 × 4, *i.e.*, 41600 digits, taken from the B reports. These tables have proved to be fairly random in character. Any page of it selected at random and the numbers in any row or column or diagonal selected at ran be taken to constitute the sample.

14·2·3. Simple Sampling. Simple sampling is random sampling in which unit of the population has an equal chance, say p, of being included in the sample that this probability is independent of the previous drawings. Thus a simple size n from a population may be identified with series of n independent tri constant probability 'p' of success for each trial.

Remark. It may be pointed out that random sampling does not necessarily imply sampling though, obviously, the converse is true. For example, if an urn contains 'a' and 'b' black balls, the probability of drawing a white ball at the first draw is $a/(a+b)$ (say) and if this ball is not replaced the probability of getting a white ball in the second $[(a-1)/(a+b-1)] = p_2 \neq p_1$, the sampling is not simple. But since in the first draw ball has the same chance, *viz.*, $a/(a+b)$, of being drawn and in the second draw white ball has the same chance, *viz.*, $(a-1)/(a+b-1)$, of being drawn, the sampling Hence in this case, the sampling, though random, is not simple. To ensure that simple. it must be done with replacement, if population is finite. However, in case population no replacement is necessary.

14·2·4. Stratified Sampling. Here the entire heterogeneous population into a number of homogeneous groups, usually termed as *strata*, which one another but each of these groups is homogeneous within itself. Then sampled at random from each of these stratum, the sample size in each strat according to the relative importance of the stratum in the population. Th

...e aggregate of the sampled units of each of the stratum, is termed as ...l sample and the technique of drawing this sample is known as *stratified* ...m. Such a sample is by far the best and can safely be considered as ...esentative of the population from which it has been drawn.

...RAMETER AND STATISTIC

...order to avoid verbal confusion with the statistical constants of the population, ...an (μ), variance σ^2, etc., which are usually referred to as *parameters*, statistical ...es computed from the sample observations alone, *e.g.*, mean ($\bar{x}$), variance (s^2), ...ve been termed by Professor R.A. Fisher as *statistics*.

...practice parameter values are not known and the estimates based on the ...e values are generally used. Thus, statistic which may be regarded as an ...e of parameter, obtained from the sample, is a function of the sample values ...may be pointed out that a statistic, as it is based on sample values and as there ...ltiple choices of the samples that can be drawn from a population, varies from ...e to sample. The determination or the characterisation of the variation (in the ...e of the statistic obtained from different samples) that may be attributed to ...e or fluctuations of sampling, is one of the fundamental problems of the ...ling theory.

...marks 1. Now onwards, μ and σ^2 will refer to the population mean and variance ...ively while the sample mean and variance will be denoted by $\bar{x}$ and s^2 respectively.

...Unbiased Estimate. A statistic $t = t(x_1, x_2, ..., x_n)$, a function of the sample values $x_1, x_2, ...$, ...an unbiased estimate of the population parameter θ, if $E(t) = \theta$. In other words, if :

$$E(Statistic) = \text{Parameter}, \qquad \ldots (14{\cdot}1)$$

...tatistic is said to be an unbiased estimate of the parameter.

...3.1. Sampling Distribution of a Statistic. If we draw a sample of size n from ...e finite population of size N, then the total number of possible samples is :

$$^{N}C_{n} = \frac{N!}{n!\,(N-n)!} = k, \text{ (say)}.$$

...for each of these k samples we can compute some statistic $t = t(x_1, x_2, ..., x_n)$, in ...ular the mean $\bar{x}$, the variance s^2, etc., as given below.

Sample Number	Statistic t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

...The set of the values of the statistic so obtained, one for each sample, constitutes ...is called the *sampling distribution* of the statistic. For example, the values $t_1, t_2, t_3, ...$ ...determine the sampling distribution of the statistic t. In other words, statistic t ...be regarded as a random variable which can take the values $t_1, t_2, t_3, ..., t_k$ and we ...tants like mean variance, skewness, kurtosis,

etc., for its distribution. For example, the mean and variance of the sampling distribution of the statistic t are given by :

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + ... + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

and $$\text{Var}(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + ... + (t_k - \bar{t})^2\right] = \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2.$$

14·3·2. Standard Error. The standard deviation of the sampling distribution of a statistic is known as its *Standard Error*, abbreviated as S.E. The standard errors of some of the well-known statistics, *for large samples*, are given below, where n is the sample size, σ^2 the population variance, and P the population proportion, and $Q = 1 - P$; n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population (s).

S. No.	Statistic	Standard Error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	Sample s.d. : s	$\sqrt{\sigma^2/2n}$
4.	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1{\cdot}36263\,\sigma/\sqrt{n}$
6.	Sample median	$1{\cdot}25331\,\sigma/\sqrt{n}$
7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient
8.	Sample moment : μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moment : μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}}\sqrt{1+\frac{2v^2}{10^4}} = \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample s.d.'s : $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions : $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Utility of Standard Error. S.E. plays a very important role in the large sa... theory and forms the basis of the testing of hypothesis. If t is any statistic, the... large samples : $Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1)$ (c.f.

$$\Rightarrow \quad Z = \frac{t - E(t)}{\text{S.E.}(t)} \sim N(0, 1), \text{ for large samples.}$$

Thus, if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than z_α (*c.f.* §14·4·5) times its S.E., the null hypothesis is rejected at α level of significance. Similarly, if

$$| t - E(t) | \leq z_\alpha \times \text{S.E.}(t),$$

the deviation is not regarded significant at 5% level of significance. In other words, the deviation, $t - E(t)$, could have arisen due to fluctuations of sampling and the data do not provide us any evidence against the null hypothesis which may, therefore, be accepted at α level of significance. [For details see § 14·4·3.]

(*i*) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the statistic.

$$S.E.(p) = \sqrt{PQ/n} \quad \text{and} \quad S.E.(\bar{x}) = \sigma/\sqrt{n}$$

In other words, the standard errors of p and $\bar{x}$ vary inversely as the square root of the sample size. Thus in order to double the precision, which amounts to reducing the standard error to one half, the sample size has to be increased four times.

(*ii*) S.E. enables us to determine the probable limits within which the population parameter may be expected to lie. For example, the probable limits for population proportion P are given by : $p \pm 3\sqrt{pq/n}$. (*c.f. Remark* § 14·7·1)

Remark. S.E. of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time, etc.

14·4. TESTS OF SIGNIFICANCE

A very important aspect of the sampling theory is the study of the *tests of significance*, which enable us to decide on the basis of the sample results, if

(*i*) the deviation between the observed sample statistic and the hypothetical parameter value, or

(*ii*) the deviation between two independent sample statistics ; is significant or might be attributed to chance or the fluctuations of sampling.

Since, for large n, almost all the distributions, *e.g.*, Binomial, Poisson, Negative binomial, Hypergeometric (*c.f.* Chapter 8), t, F (Chapter 16), Chi-square (Chapter 15), can be approximated very closely by a normal probability curve, we use the *Normal Test of Significance* (*c. f.* § 14·7) for large samples. Some of the well-known tests of significance for studying such differences for small samples are *t-test*, *F-test* and Fisher's *z-transformation*.

14·4·1. Null and Alternative Hypotheses. The technique of randomisation used for the selection of sample units makes the test of significance valid for us. For applying the test of significance we first set up a hypothesis—a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called *null hypothesis* and is usually denoted by H_0. According to *Prof. R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.*

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

Having set up the null hypothesis we compute the probability ... deviation between the observed sample statistic and the hypothetical param... might have occurred due to fluctuations of sampling. If the deviation come... significant (as measured by a test of significance) null hypothesis is refuted ... at the particular level of significance adopted (*c.f.* § 14·4·3) and if the devia... significant, null hypothesis may be retained at that level.

Any hypothesis which is complementary to the null hypothesis is ... *alternative hypothesis*, usually denoted by H_1. For example, if we want to t... hypothesis that the population has a specified mean μ_0, (say), *i.e.*, $H_0 : \mu = \mu_0$, ... alternative hypothesis could be :

(*i*) $H_1 : \mu \neq \mu_0$ (*i.e.*, $\mu > \mu_0$ or $\mu < \mu_0$) (*ii*) $H_1 : \mu > \mu_0$, (iii) $H_1 : \mu < \mu_0$

The alternative hypothesis in (*i*) is known as a *two-tailed alternative* ... alternatives in (*ii*) and (*iii*) are known as *right-tailed* and *left-tailed* alt... respectively. The setting of alternative hypothesis is very important since it ... to decide whether we have to use a single-tailed (right or left) or two-tailed ... 14·4·4).

14·4·2. Errors in Sampling. The main objective in sampling theory is ... valid inferences about the population parameters on the basis of the sample ... practice we decide to accept or reject the lot after examining a sample from it, ... we are liable to commit the following two types of errors :

Type I Error : *Reject H_0 when it is true.*

Type II Error : *Accept H_0 when it is wrong, i.e., accept H_0 when H_1 is true.*

If we write P {Reject H_0 when it is true} = P {Reject H_0 | H_0} = α

and P {Accept H_0 when it is wrong} = P {Accept H_0 | H_1} = β

then α and β are called the *sizes of type I error and type II error*, respectively.

In practice, type I error amounts to rejecting a lot when it is good ... error may be regarded as accepting the lot when it is bad.

Thus P {Reject a lot when it is good} = α

and P {Accept a lot when it is bad} = β

where α and β are referred to as *Producer's risk* and *Consumer's risk* respectively.

14·4·3. Critical Region and Level of Significance. A region (correspondi... to a statistic t) in the sample space S which amounts to rejection of H_0 is ... *critical region of rejection*. If ω is the critical region and if $t = t(x_1, x_2, \ldots, x_n)$ is ... of the statistic based on a random sample of size n, then

$$P(t \in \omega \mid H_0) = \alpha, \quad P(t \in \bar{\omega} \mid H_1) = \beta$$

where $\bar{\omega}$, the complementary set of ω, is called the *acceptance region*.

We have $\omega \cup \bar{\omega} = S$ and $\omega \cap \bar{\omega} = \phi$

The probability 'α' that a random value of the statistic t belongs to the ... region is known as the *level of significance*. In other words, level of significa... size of the type I error (or the maximum producer's risk). The levels of sig... usually employed in testing of hypothesis are 5% and 1%. The level of signi... always fixed in advance before collecting the sample information.

14·4·4. One-tailed and Two-tailed Tests. In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

A test of any statistical hypothesis where the alternatives hypothesis is one tailed (right-tailed or left-tailed) is called a *one-tailed test*. For example, a test for testing the mean of a population $H_0 : \mu = \mu_0$ against the alternative hypothesis :

$H_1 : \mu > \mu_0$ (*Right-tailed*) or $H_1 : \mu < \mu_0$ (*Left-tailed*), is a *single tailed test.*

In the right-tailed test ($H_1 : \mu > \mu_0$), the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left-tailed test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the distribution.

A test of statistical hypothesis where the alternative hypothesis is two-tailed such as : $H_0 : \mu = \mu_0$, against the alternative hypothesis $H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$), is known as *two tailed test* and in such a case the critical region is given by the portion of the area lying in both tails of the probability curve of the test statistic.

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed, we apply two-tailed test and if alternative hypothesis is one-tailed, we apply one tailed test.

For example, suppose that there are two population brands of bulbs, one manufactured by standard process (with mean life μ_1) and the other manufactured by some new technique (with mean life μ_2). If we want to test if the bulbs differ significantly, then our null hypothesis is $H_0 : \mu_1 = \mu_2$ and alternative will be $H_1 : \mu_1 \neq \mu_2$, thus giving us two-tailed test. However, if we want to test if the bulbs produced by new process have higher average life than those produced by standard process, then we have $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$, thus giving us a left-tailed test. Similarly, for testing if the product of new process is inferior to that of standard process, then we have : $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$, thus giving us a right-tailed test. Thus, the decision about applying a two-tailed test or a single-tailed (right or left) test will depend on the problem under study.

14·4·5. Critical Values or Significant Values. The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the *critical value* or *significant value*. It depends upon :

(i) The level of significance used, and

(ii) The alternative hypothesis, whether it is two-tailed or single-tailed.

As has been pointed out earlier, for large samples, the standardised variable corresponding to the statistic t, *viz.*,

$$Z = \frac{t - E(t)}{S.E(t)} \sim N(0, 1), \qquad \ldots (*)$$

is asymptotically as $n \to \infty$. The value of Z given by (*) under the null hypothesis is known as *test statistic*. The critical value of the test statistic at level of significance α for two-tailed test is given by z_α, where z_α is determined by the equation :

$$P(|Z| > z_\alpha) = \alpha \qquad \ldots (14{\cdot}2c)$$

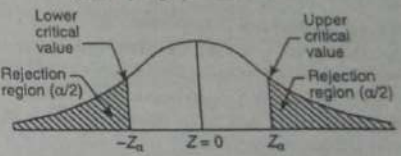
z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is a symmetrical curve, from (14·2c), we get

$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \quad \Rightarrow \quad P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha \qquad \text{[By symmetry]}$$

$$P(Z > z_\alpha) = \alpha/2$$

In other words, the area of each tail is $\alpha/2$. Thus z_α is the value such that area to the right of z_α is $\alpha/2$ and to the left of $(-z_\alpha)$ is $\alpha/2$, as shown in the following diagram :

TWO-TAILED TEST
(*Level of Significance 'α'*)

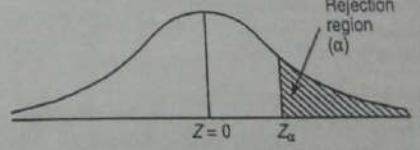


In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $(-z_\alpha)$ is α (See diagrams below), *i.e.*,

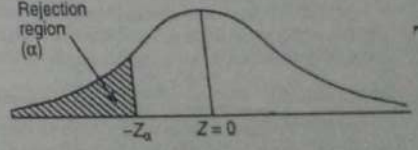
For Right-tailed Test : $P(Z > z_\alpha) = \alpha$ ... (14·2d)

For Left-tailed Test : $P(Z < -z_\alpha) = \alpha$ ... (14·2e)

RIGHT-TAILED TEST
(*Level of Significance 'α'*)



LEFT-TAILED TEST
(*Level of Significance 'α'*)



Thus *the significant or critical value of Z for a single-tailed test (left or right) at level of significance 'α' is same as the critical value of Z for a two-tailed test at level of significance '2α'.*

We give below, the critical values of Z at commonly used levels of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (14·2c), (14·2d), (14·2e), on using the Normal Probability Tables as explained in § 14·6.

Critical value (z_α)	*Level of significance* (α) 1%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2{\cdot}58$	$\lvert Z_\alpha \rvert = 1{\cdot}96$	$\lvert Z_\alpha \rvert = 1{\cdot}645$
Right-tailed test	$Z_\alpha = 2{\cdot}33$	$Z_\alpha = 1{\cdot}645$	$Z_\alpha = 1{\cdot}28$
Left-tailed test	$Z_\alpha = -2{\cdot}33$	$Z_\alpha = -1{\cdot}645$	$Z_\alpha = -1{\cdot}28$

Remark. If n is small, then the sampling distribution of the test statistic Z will not be normal and in the that case we can't use the above significant values which have been obtained from normal probability curves. In this case, *viz.*, n small (usually less than 30), we use t

significant values based on the exact sampling distribution of the statistic Z, (defined in (*), § 14·4·3), which turns out to be t, F, or χ^2 (see Chapters 14 and 16). These significant values have been tabulated for different values of n and α and are given at the end of Chapters 15 and 16.

14·5. PROCEDURE FOR TESTING OF HYPOTHESIS

We now summarise below the various steps in testing of a statistical hypothesis in a systematic manner:

1. *Null Hypothesis.* Set up the Null Hypothesis H_0.

2. *Alternative Hypothesis.* Set up the Alternative Hypothesis H_1. This will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3. *Level of Significance.* Choose the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn, *i.e.*, α is fixed in advance.

4. *Test Statistic (or Test Criterion).* Compute the test statistic :

$$Z = \frac{t - E(t)}{S.E.(t)}, \text{ under } H_0$$

5. *Conclusion.* We compare the computed value of Z in *step* 4 with the significant value (tabulated value) z_α at the given level of significance, 'α'.

If $|Z| < z_\alpha$, *i.e.*, if the calculated value of Z (in modulus value) is less than z_α we say it is not significant. By this we mean that the difference $t - E(t)$ is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may, therefore, be accepted.

If $|Z| > z_\alpha$ *i.e.*, if the computed value of test statistic is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance α, *i.e.*, with confidence coefficient $(1 - \alpha)$.

14·6. TESTS OF SIGNIFICANCE FOR LARGE SAMPLES

In this section, we will discuss the tests of significance when samples are large. We have seen that for large values of n, the number of trials, almost all the distributions, *e.g.*, binomial, Poisson, negative binomial, etc., are very closely approximated by normal distribution. Thus in this case we apply the *normal test*, which is based upon the following fundamental property (*area property*) of the normal probability curve.

If $X \sim N(\mu, \sigma^2)$, then $$Z = \frac{X - \mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0, 1)$$

Thus from the normal probability tables, we have

$$P(-3 \le Z \le 3) \simeq 0.9973, \quad i.e., \quad P(|Z| \le 3) = 0.9973$$

$$\Rightarrow \quad P(|Z| > 3) = 1 - P(|Z| \le 3) = 0.0027 \qquad \ldots (14{\cdot}3)$$

i.e., in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability tables, we get

$$P(-1.96 \le Z \le 1.96) = 0.95, \quad i.e., \quad P(|Z| \le 1.96) = 0.95$$

$$\Rightarrow \quad P(|Z| > 1.96) = 1 - 0.95 = 0.05 \qquad \ldots (14{\cdot}3a)$$

and $$P(|Z| < 2.58) = 0.99 \Rightarrow P(|Z| > 2.58) = 0.01 \qquad \ldots (14{\cdot}3b)$$

Thus the significant values of Z at 5% and 1% levels of significance for a two-tailed test are 1·96 and 2·58 respectively.

Thus the steps to be used in the normal test are as follows :

(*i*) Compute the test statistic Z under H_0.

(*ii*) If $|Z| > 3$, H_0 is always rejected.

(*iii*) If $|Z| \le 3$, we test its significance at certain level of significance, usually 5% and sometimes at 1% level of significance. Thus, for a two-tailed test if $|Z| > 1.96$, H_0 is rejected at 5% level of significance.

Similarly if $|Z| > 2.58$, H_0 is contradicted at 1% level of significance and if $|Z| \le 2.58$, H_0 may be accepted at 1% level of significance.

From the normal probability tables, we have :

$$P(Z > 1.645) = 0.5 - P(0 \le Z \le 1.645) = 0.5 - 0.45 = 0.05$$

$$P(Z > 2.33) = 0.5 - P(0 \le Z \le 2.33) = 0.5 - 0.49 = 0.01$$

Hence for a single-tail test (Right-tail or Left-tail) we compare the computed value of $|Z|$ with 1·645 (at 5% level) and 2·33 (at 1% level) and accept or reject accordingly.

Important Remark. In the theoretical discussion that follows in the next section, samples under consideration are supposed to be large. For practical purposes, samples may be regarded as large if $n > 30$.

14·7. SAMPLING OF ATTRIBUTES

Here we shall consider sampling from a population which is divided into two mutually exclusive and collectively exhaustive classes—one class possessing a particular attribute, say A, and the other class not possessing that attribute, $\bar{A}$. We note down the number of persons in the sample of size n, possessing that attribute. The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of n observations is identified with a series of n independent Bernoulli trials with constant probability P of success in each trial. Then the probability of x successes in n trials, as given by the binomial probability distribution is : $$p(x) = {}^nC_x P^x Q^{n-x}; \; x = 0, 1, 2, \ldots, n$$

14·7·1. Test of Significance for Single Proportion. If X is the number of successes in n independent trials with constant probability P of success for each trial, then

$$E(X) = nP \quad \text{and} \quad V(X) = nPQ, \quad \text{where} \quad Q = 1 - P, \text{ is the probability of failure.}$$

It has been proved that for large n, the binomial distribution tends to normal distribution. Hence for large n, $X \sim N(nP, nPQ)$, *i.e.*,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1)$$

and we can apply the normal test.

Remarks 1. In a sample of size n, let X be the number of persons possessing the attribute. Then

Observed proportion of successes $= X/n = p$, (say).

$$\therefore \quad E(p) = E(X/n) = \frac{1}{n} E(X) = \frac{1}{n} nP = P$$

Thus the sample proportion 'p' gives an unbiased estimate of the population proportion P.

Also $V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n} \Rightarrow$ S.E. $(p) = \sqrt{\frac{PQ}{n}}$ ... (14-4b)

Since X and consequently X/n is asymptotically normal for large n, the normal test for the proportion of successes becomes :

$$Z = \frac{p - E(p)}{S.E.(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1) \quad \text{... (14-4c)}$$

2. If we have sampling from a finite population of size N, then

$$S.E.(p) = \sqrt{\left(\frac{N-n}{N-1}\right) \cdot \frac{PQ}{n}} \quad \text{... (14-4d)}$$

3. Since the probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$, the probable limits for the observed proportion of successes are :

$$E(p) \pm 3\,S.E.(p), \quad i.e., \quad P \pm 3\sqrt{PQ/n}.$$

4. If P is not known then taking p (the sample proportion) as an estimate of P, the probable limits for the proportion in the population are : $p \pm 3\sqrt{pq/n}$. ... (14-4e)

However, the limits for P at level of significance α are given by : $p \pm z_\alpha \sqrt{pq/n}$, ... (14-4f)

where z_α is the significant value of Z at level of significance α.

In particular : 95% confidence limits for P are given by : $p \pm 1{\cdot}96\sqrt{pq/n}$, ... (14-4g)

99% confidence limits for P are given by : $p \pm 2{\cdot}58\sqrt{pq/n}$. ... (14-4h)

Example 14·1. *A die is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the die cannot be regarded as an unbiased one and find the limits between which the probability of a throw of 3 or 4 lies.*

Solution. If the coming of 3 or 4 is called a success, then in usual notations :

$n = 9{,}000$; X = Number of successes = 3,240

Under the null hypothesis (H_0) that the die is an unbiased one, we get

P = Probability of success = Probability of getting a 3 or 4 $= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

Alternative hypothesis, $H_1 : p \neq \frac{1}{3}$, (i.e., die is biased).

We have $Z = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1)$, since n is large.

Now $$Z = \frac{3240 - 9000 \times (1/3)}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = \frac{240}{44{\cdot}73} = 5{\cdot}36$$

Since $|Z| > 3$, H_0 is rejected and we conclude that the die is almost certainly biased.

Since die is not unbiased, $P \neq \frac{1}{3}$. The probable limits for 'P' are given by :

$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm 3\sqrt{pq/n}$, where $\hat{P} = p = \frac{3{,}240}{9{,}000} = 0{\cdot}36$ and $\hat{Q} = q = 1 - p = 0{\cdot}64$.

Probable limits for population proportion of successes may be taken as :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}36 \pm 3\sqrt{\frac{0{\cdot}36 \times 0{\cdot}64}{9000}} = 0{\cdot}36 \pm 3 \times \frac{0{\cdot}6 \times 0{\cdot}8}{30\sqrt{10}} = 0{\cdot}345 \text{ and } 0{\cdot}375.$$

Hence the probability of getting 3 or 4 almost certainly lies between 0·345 and

Example 14·2. *A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the proportion of bad ones in a sample of this size is 0·015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8·5 and 17·5.*

Solution. Here we are given : $n = 500$

X = Number of bad pineapples in the sample = 65

p = Proportion of bad pineapples in the sample $= \frac{65}{500} = 0{\cdot}13 \Rightarrow q = 1 - p = 0{\cdot}87$

Since P, the proportion of bad pineapples in the consignment is not known, we may take (as in the last example) : $\hat{P} = p = 0{\cdot}13$, $\hat{Q} = q = 0{\cdot}87$.

S.E. of proportion $= \sqrt{\hat{P}\hat{Q}/n} = \sqrt{0{\cdot}13 \times 0{\cdot}87/500} = 0{\cdot}015$

Thus, the limits for the proportion of bad pineapples in the consignment are :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}130 \pm 3 \times 0{\cdot}015 = 0{\cdot}130 \pm 0{\cdot}045 = (0{\cdot}085, 0{\cdot}175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8·5 and 17·5.

Example 14·3. *A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage of bad apples in the consignment.*

Solution. We have :

p = Proportion of bad apples in the sample $= \frac{60}{500} = 0{\cdot}12$

Since significant value of Z at 98% confidence coefficient (level of significance) is 2·33, [from Normal Tables], 98% confidence limits for population proportion are

$$p \pm 2{\cdot}33\sqrt{pq/n} = 0{\cdot}12 \pm 2{\cdot}33\sqrt{0{\cdot}12 \times 0{\cdot}88/500} = 0{\cdot}12 \pm 2{\cdot}33 \times \sqrt{0{\cdot}0002112}$$

$$= 0{\cdot}12 \pm 2{\cdot}33 \times 0{\cdot}01453 = (0{\cdot}08615, 0{\cdot}15385)$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8·61, 15·38).

Example 14·4. *In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance ?*

Solution. In the usual notations, we are given : $n = 1{,}000$

X = Number of rice eaters = 540

$\therefore$ p = Sample proportion of rice eaters $= \frac{X}{n} = \frac{540}{1000} = 0{\cdot}54$

Null Hypothesis, H_0 : Both rice and wheat are equally popular in the State.

P = Population proportion of rice eaters in Maharashtra = 0·5 $\Rightarrow Q$ =

Alternative Hypothesis, $H_1 : P \neq 0{\cdot}5$ (two-tailed alternative)

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since } n \text{ is large).}$$

Now $$Z = \frac{0.54 - 0.50}{\sqrt{0.5 \times 0.5/1000}} = \frac{0.04}{0.0158} = 2.532$$

Conclusion. The significant or critical value of Z at 1% level of significance for two-tailed test is 2·58. Since computed $Z = 2.532$ is less than 2·58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

Example 14·5. *Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)*

Solution. In the usual notations, we are given : $n = 20$.

X = Number of persons who survived after attack by a disease = 18

p = Proportion of persons survived in the sample $= \frac{18}{20} = 0.90$

Null Hypothesis, $H_0 : P = 0.85$, *i.e.,* the proportion of persons survived after attack by a disease in the lot is 85%.

Alternative Hypothesis, $H_1 : P > 0.85$ (Right-tailed alternative).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1), \text{ (since sample is large).}$$

Now $$Z = \frac{0.90 - 0.85}{\sqrt{0.85 \times 0.15/20}} = \frac{0.05}{0.079} = 0.633$$

Conclusion. Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of Z. The significant value of Z at 5% level of significance for right-tailed test is + 1·645. Since computed value of $Z = 0.633$ is less than 1·645, it is not significant and we may accept the null hypothesis at 5% level of significance.

Example 14·6. *Work sampling studies are conducted to find the utilization of a machine. Out of 200 observations made, only 40 observations indicated the machine to be idle. Find the number of observations to be made in order to satisfy 95% confidence to state the utilization of machine with expected accuracy of ± 5%.*

Solution. The sample size for n estimating the population proportion P with confidence coefficient $(1 - \alpha) = 0.95$ is given by the equation :

$$P_r[\,|p - P| \le 1.96\sqrt{PQ/n}\,] = 0.95 \quad \ldots (*)$$

We want $$P_r\,[\,|p - P| < 0.05] = 0.95. \quad \ldots (**)$$

From (*) and (**), we obtain $1.96\sqrt{PQ/n} = 0.05 \Rightarrow n = \frac{PQ(1.96)^2}{(0.05)^2}$

Since P is not known, we may use its sample estimate $\hat{P} = p = \frac{40}{200} = 0.2$.

$$\therefore \quad n = \frac{0.2 \times (1 - 0.2) \times 3.8416}{0.0025} = \frac{0.6147}{0.0025} = 245.88 \simeq 246.$$

14·7·2. Test of Significance for Differ

compare two distinct populations with respect to the prevalence of [illegible], say A, among their members. Let X_1, X_2 be the number of persons possessing given attribute A in random samples of sizes n_1 and n_2 from the respectively. Then sample proportions are given by : $p_1 = X_1/n_1$ and

If P_1 and P_2 are population proportions, then

$$E(p_1) = P_1, E(p_2) = P_2$$

and $$V(p_1) = \frac{P_1 Q_1}{n_1} \quad \text{and} \quad V(p_2) = \frac{P_2 Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are independently and asymptotically distributed, $(p_1 - p_2)$ is also normally distributed. Then the standard variate corresponding to the difference $(p_1 - p_2)$ is given by :

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the *null hypothesis,* $H_0 : P_1 = P_2$, i.e., there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(p_1) = E(p_2) = P_1 - P_2 = 0$$

Also $$V(p_1 - p_2) = V(p_1) + V(p_2),$$

the covariance term $\text{Cov}(p_1, p_2)$ vanishes, since samples are independent.

$$\Rightarrow \quad V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \quad [\because \text{under } H_0 : P_1 = P_2 = P]$$

Hence, under $H_0 : P_1 = P_2$, the test statistic for the difference becomes :

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

In general, we do not have any information as to the proportion in the populations from which the samples have been taken. Under $H_0 : P_1 = P_2 = P$, an unbiased estimate of the population proportion P, based on both the samples is given by :

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

The estimate is unbiased, since

$$E(\hat{P}) = \frac{1}{n_1 + n_2} E[n_1 p_1 + n_2 p_2] = \frac{1}{n_1 + n_2}[n_1 E(p_1) + n_2 E(p_2)]$$

$$= \frac{1}{n_1 + n_2}(n_1 P_1 + n_2 P_2) = P \quad [\because P_1 = P_2 = P]$$

Thus (14·5) along with (14·5a) gives the required test statistic.

Remarks 1. Suppose we want to test the significance of the difference between p, where $p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$ gives a pooled estimate of the population proportion in both the samples. We have $V(p_1 - p) = V(p_1) + V(p) - 2\,\text{Cov}(p_1, p)$.

Since p_1 and p are not independent, $\text{Cov}(p_1, p) \neq 0$.

$$\text{Cov}(p_1, p) = E[\{p_1 - E(p_1)\}\{p - E(p)\}]$$

$$= E\left[\{p_1 - E(p_1)\}\left\{\frac{1}{n_1+n_2}\{n_1p_1 + n_2p_2 - E(n_1p_1 + n_2p_2)\}\right\}\right]$$

$$= \frac{1}{n_1+n_2} E\left[\{p_1 - E(p_1)\}\{n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2))\}\right]$$

$$= \frac{1}{n_1+n_2}\left[n_1 E\{p_1 - E(p_1)\}^2 + n_2 E\{(p_1 - E(p_1))(p_2 - E(p_2))\}\right]$$

$$= \frac{1}{n_1+n_2}\left[n_1 V(p_1) + n_2 \text{Cov}(p_1, p_2)\right] = \frac{1}{n_1+n_2} n_1 V(p_1), \quad [\because \text{Cov}(p_1, p_2) = 0]$$

$$= \frac{n_1}{n_1+n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1+n_2}.$$

$$\text{Var}(p) = \text{Var}\left[\frac{n_1p_1 + n_2p_2}{n_1+n_2}\right] = \frac{1}{(n_1+n_2)^2}\text{Var}(n_1p_1 + n_2p_2)$$

$$= \frac{1}{(n_1+n_2)^2}\left[n_1^2 \text{Var}(p_1) + n_2^2 \text{Var}(p_2)\right],$$

since cross term vanishes since p_1 and p_2 are independent.

$$\text{Var}(p) = \frac{1}{(n_1+n_2)^2}\left(n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2}\right) = \frac{pq}{n_1+n_2}$$

Substituting (*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1+n_2} - 2\frac{pq}{n_1+n_2} = pq\left[\frac{n_2}{n_1(n_1+n_2)}\right]$$

Also $E(P_1 - P) = E(P_1) - E(P) = P - P = 0$

Thus, the test statistic in this case becomes :

$$\left(\therefore Z = \frac{(p_1 - p) - E(p_1 - p)}{\text{S.E.}(p_1 - p)} = \frac{p_1 - p}{\sqrt{\left\{\frac{n_2}{(n_1+n_2)}\frac{pq}{n_1}\right\}}} \sim N(0, 1)\right) \quad \ldots (14\cdot5b)$$

Suppose the population proportions P_1 and P_2 are given to be distinctly different, *i.e.*, and we want to test if the difference $(P_1 - P_2)$ in population proportions is likely to be in simple samples of sizes n_1 and n_2 from the two populations respectively.

We have seen that in the usual notations,

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E.}(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}} \sim N(0, 1)$$

ere sample proportions are not given. If we set up the *null hypothesis* $H_0 : p_1 = p_2$, *i.e.*, the es will not reveal the difference in the population proportions or in other words the ence in population proportions is likely to be hidden in sampling, the test statistic es :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}} \sim N(0, 1) \quad \ldots (14\cdot5c)$$

Example 14·7. Random samples of 400 men and 600 women were asked whether they like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are against that they are not, at 5% level.

Solution. *Null Hypothesis* $H_0 : P_1 = P_2 = P$ (say), *i.e.*, there is no significant difference between the opinions of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

We are given :

$n_1 = 400$, X_1 = Number of men favouring the proposal = 200

$n_2 = 600$, X_2 = Number of women favouring the proposal = 325

$\therefore$ p_1 = Proportion of men favouring the proposal in the sample $= \frac{X_1}{n_1} = \frac{200}{400} = 0.5$

p_2 = Proportion of women favouring the proposal in the sample $= \frac{X_2}{n_2} = \frac{325}{600} = 0.541$

Test Statistic. Since samples are large, the test statistic under the Null Hypothesis, H_0 is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1), \text{ where}$$

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1+n_2} = \frac{X_1 + X_2}{n_1+n_2} = \frac{200 + 325}{400 + 600} = 0.525 \quad \Rightarrow \quad \hat{Q} = 1 - \hat{P} = 1 - 0.525 = 0.475$$

$$\therefore Z = \frac{0.500 - 0.541}{\sqrt{0.525 \times 0.475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = \frac{-0.041}{\sqrt{0.001039}} = \frac{-0.041}{0.0323} = -1.269$$

Conclusion. Since $|Z| = 1.269$ which is less than 1·96, it is not significant at 5% level of significance. Hence H_0 may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

Example 14·8. *In a large city A, 20 per cent of a random sample of 900 school children had defective eye-sight. In other large city B, 15 per cent of random sample of 1,600 children had the same defect. Is this difference between the two proportions significant ? Obtain 95% confidence limits for the difference in the population proportions.*

Solution. In usual notations : $n_1 = 900$, $p_1 = 20\% = 0.20$, $n_2 = 1600$, $p_2 = 15\% = 0.15$.

Null hypothesis, $H_0 : P_1 = P_2$.

Alternative hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} \sim N(0, 1), \text{ (since the samples are large.)}$$

where $\hat{P} = \frac{900 \times 0.20 + 1600 \times 0.15}{900 + 1600} = 0.168 \quad \Rightarrow \quad \hat{Q} = 1 - \hat{P} = 0.832$

$$\text{S.E.}(p_1 - p_2) = \sqrt{\left\{0.168 \times 0.832\left(\frac{1}{900} + \frac{1}{1600}\right)\right\}} = \sqrt{0.0002427} = 0.0156.$$

$$\therefore \quad Z = \frac{0.20 - 0.15}{0.0156} = 3.21$$

Conclusion. Since the calculated value of Z is greater than 1·96, it is significant at 5% level. We, therefore, reject the null hypothesis H_0 and conclude that the difference between the two proportions is significant.

The 95% confidence limits for the difference $P_1 - P_2$ are :

$(p_1 - p_2) \pm 1{\cdot}96$ S.E. of $(p_1 - p_2)$,

where $$\text{S.E. of } (p_1 - p_2) = \sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)} \simeq \sqrt{\left(\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\right)}$$

$$\simeq \sqrt{\left(\frac{0{\cdot}20 \times 0{\cdot}80}{900} + \frac{0{\cdot}15 \times 0{\cdot}85}{1600}\right)} = 0{\cdot}016.$$

Hence, the 95% confidence limits for $P_1 - P_2$ are :

$(0{\cdot}20 - 0{\cdot}15) \pm 1{\cdot}96\ (0{\cdot}016) = 0{\cdot}05 \pm 0{\cdot}031 = (0{\cdot}019 \text{ and } 0{\cdot}081)$.

Example 14·9. *A company has the head office at Kolkata and a branch at Mumbai. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Kolkata, 62% favoured the new plan. At Mumbai out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level ?*

Solution. In the usual notations, we are given :

$$n_1 = 500,\ p_1 = 0{\cdot}62 \quad \text{and} \quad n_2 = 400,\ p_2 = 1 - 0{\cdot}41 = 0{\cdot}59$$

Null hypothesis, $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference between the two groups in their attitude towards the new plan.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{p_1 - p_2}{\text{S.E. } (p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N\,(0, 1), \text{ where}$$

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{500 \times 0{\cdot}62 + 400 \times 0{\cdot}59}{500 + 400} = 0{\cdot}607 \Rightarrow \hat{Q} = 1 - \hat{P} = 0{\cdot}393$$

$$\therefore\ Z = \frac{0{\cdot}62 - 0{\cdot}59}{\sqrt{\left\{0{\cdot}607 \times 0{\cdot}393 \times \left(\frac{1}{500} + \frac{1}{400}\right)\right\}}} = \frac{0{\cdot}03}{\sqrt{0{\cdot}00107}} = \frac{0{\cdot}03}{0{\cdot}0327} = 0{\cdot}917.$$

Critical region. At 5% level of significance, the critical value of Z for a two-tailed test is 1·96. Thus the critical region consists of all values of $Z \geq 1{\cdot}96$ or $Z \leq -1{\cdot}96$.

Conclusion. Since the calculated value of $|Z| = 0{\cdot}917$ is less than the critical value of Z (1·96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we may conclude that there is no significant difference between the two groups in their attitude towards the new plan.

Example 14·10. *Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in duty, 800 persons were found to be tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty.*

Solution. In the usual notations, we have $n_1 = 1{,}000$; $n_2 = 1{,}200$.

p_1 = Sample proportion of tea drinkers before increase in excise duty

p_2 = Sample proportion of tea drinkers after increase in excise duty

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference in consumption of tea before and after the increase in excise duty.

Alternative Hypothesis, $H_1 : P_1 > P_2$ (Right-tailed alternative).

Test Statistic. Under the null hypothesis, the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N\,(0, 1)$$

where $$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{16}{22}, \quad \text{and} \quad \hat{Q} = 1 - \hat{P} = \frac{6}{22}$$

$$\therefore\ Z = \frac{0{\cdot}80 - 0{\cdot}67}{\sqrt{\left\{\frac{16}{22} \times \frac{6}{22} \times \left(\frac{1}{1000} + \frac{1}{1200}\right)\right\}}} = \frac{0{\cdot}13}{0{\cdot}019} = 6{\cdot}842$$

Conclusion. Since Z is much greater than 1·645 as well as 2·33 (since the test is right-tailed), it is highly significant at both 5% and 1% levels of significance. We reject the null hypothesis H_0 and conclude that there is a significant decrease in consumption of tea after increase in the excise duty.

Example 14·11. *A cigarette manufacturing firm claims that its brand A outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B, test the difference is a valid claim. (Use 5% level of significance.)*

Solution. In usual notations, we are given :

$$n_1 = 200,\ X_1 = 42 \Rightarrow p_1 = \frac{X_1}{n_1} = \frac{42}{200} = 0{\cdot}21$$

$$n_2 = 100,\ X_2 = 18 \Rightarrow p_2 = \frac{X_2}{n_2} = \frac{18}{100} = 0{\cdot}18$$

We set up the *Null Hypothesis* that 8% difference in the sale of the cigarettes is a valid claim, *i.e.,* $H_0 : P_1 - P_2 = 0{\cdot}08$.

Alternative Hypothesis : $H_1 : P_1 - P_2 \neq 0{\cdot}08$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\left\{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N\,(0, 1)$$

where $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{42 + 18}{200 + 100} = \frac{60}{300} = 0.20 \Rightarrow \hat{Q} = 1 - \hat{P} = 0.80$

$$Z = \frac{(0.21 - 0.18) - 0.08}{\sqrt{\left\{0.2 \times 0.8\left(\frac{1}{200} + \frac{1}{100}\right)\right\}}} = \frac{-0.05}{\sqrt{0.16 \times 0.015}} = \frac{-0.05}{0.04899} = -1.02$$

Since $|Z| = 1.02 < 1.96$, it is not significant at 5% level of significance. Hence ...hypothesis may be retained at 5% level of significance and we may conclude that ...ference of 8% in the sale of two brands of cigarettes is a valid claim by the firm.

Example 14·12. *On the basis of their total scores, 200 candidates of a civil service ...mination are divided into two groups, the upper 30 per cent and the remaining 70 per cent. ...ong the first question of this examination. Among the first group, 40 had the correct ...wer, whereas among the second group, 80 had the correct answer. On the basis of these ...ults, can one conclude that the first question is no good at discriminating ability of the type ...ng examined here ?*

Solution. Here, we have

n = Total number of candidates = 200

n_1 = The number of candidates in the upper 30% group = $\frac{30}{100} \times 200 = 60$

n_2 = The number of candidates in the remaining 70% group = $\frac{70}{100} \times 200 = 140$.

X_1 = The number of candidates, with correct answer in the first group = 40

X_2 = The number of candidates, with correct answer in the second group = 80

$$p_1 = \frac{X_1}{n_1} = \frac{40}{60} = 0.6666 \quad \text{and} \quad p_2 = \frac{X_2}{n_2} = \frac{80}{140} = 0.5714$$

∴ *Null Hypothesis, H_0* : There is no significance difference in the sample proportions, ...$P_1 = P_2$. In other words, the first question is not good at discriminating the ability ...of the type being examined here.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$.

Test Statistic. Under H_0 the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

where $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{40 + 80}{60 + 140} = 0.6, \quad \hat{Q} = 1 - \hat{P} = 0.4$

$$Z = \frac{0.6666 - 0.5714}{\sqrt{\left\{0.6 \times 0.4\left(\frac{1}{60} + \frac{1}{140}\right)\right\}}} = \frac{0.0953}{0.0756} = 1.258$$

Conclusion. Since $|Z| < 1.96$, the data are consistent with the null hypothesis at 5% level of significance. Hence we conclude that the first question is not good enough to distinguish between the ability of the two groups of candidates.

Example 14·13. *In a year there are 956 births in a town A, of which 52·5% were males, while in towns A and B combined, this proportion in a total of 1,406 births was 0·496. Is there ...in the proportion of male births in the two towns ?*

Solution. In usual notations, we are given :

$n_1 = 956$, $n_1 + n_2 = 1,406$ or $n_2 = 1,406 - 956 = 450$.

p_1 = Proportion of males in the sample of town $A = 0.525$.

Let p_2 be the proportion of males in the sample (of size n_2) of town B. Then

$\hat{P}$ = Proportion of males in both the samples combined = $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0.496$ (Given)

$$\therefore \quad \frac{956 \times 0.525 + 450 \times p_2}{1,406} = 0.496 \quad \Rightarrow \quad p_2 = 0.434 \qquad \text{(On simplification)}$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the proportion of male births in the two towns A and B.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

where $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0.496 \Rightarrow \hat{Q} = 1 - \hat{P} = 0.504$

$$\therefore \quad Z = \frac{0.525 - 0.434}{\sqrt{\left\{0.496 \times 0.504\left(\frac{1}{956} + \frac{1}{450}\right)\right\}}} = \frac{0.091}{0.027} = 3.368$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected, *i.e.*, the data ar... inconsistent with the hypothesis $P_1 = P_2$ and we conclude that there is significa... difference in the proportion of male births in the towns A and B.

Example 14·14. *In two large populations, there are 30 and 25 per cent respectivel... blue-eyed people. Is this difference likely to be hidden in samples of 1,200 and ... respectively from the two populations ?*

Solution. Here, we are given : $n_1 = 1,200$, $n_2 = 900$.

P_1 = Proportion of blue-eyed people in the first population = 30% = 0...

P_2 = Proportion of blue-eyed people in the second population = 25% = 0...

$\therefore \quad Q_1 = 1 - P_1 = 0.70$ and $Q_2 = 1 - P_2 = 0.75$

We set up the *null hypothesis*, H_0 that $p_1 = p_2$, *i.e.*, the sample proportions ar... In other words, the difference in population proportions is likely to be hi... sampling.

Test Statistic. Under $H_0 : p_1 = p_2$, the test statistic is :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\left(\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since sample...}$$

$$\therefore \quad |Z| = \frac{0.30 - 0.25}{\sqrt{\left(\frac{0.3 \times 0.7}{1,200} + \frac{0.25 \times 0.75}{900}\right)}} = \frac{0.05}{0.0195} = 2.56$$

Conclusion. Since $|Z| > 1.96$, the null hypothesis ($p_1 = p_2$) is refuted at 5% level of significance and we conclude that the difference in population proportions is unlikely to be hidden in sampling. In other words, the samples will reveal the difference in the population proportions.

Example 14.15. *In a random sample of 400 students of the university teaching departments, it was found that 300 students failed in the examination. In another random sample of 500 students of the affiliated colleges, the number of failures in the same examination was found to be 300. Find out whether the proportion of failures in the university teaching departments is significantly greater than the proportion of failures in the university teaching departments and affiliated colleges taken together.*

Solution. Here we are given : $n_1 = 400$, $n_2 = 500$, $p_1 = \frac{300}{400} = 0.75$, $p_2 = \frac{300}{500} = 0.60$

$\therefore \quad q_1 = 1 - p_1 = 1 - 0.75 = 0.25$ and $q_2 = 1 - p_2 = 1 - 0.60 = 0.40$

Here we set up the *null hypothesis*, H_0 that p_1 and $p = \hat{P}$, where $\hat{P}$ is the pooled estimate, *i.e.*, proportion of failures in the university teaching departments and affiliated colleges taken together, do not differ singificantly.

$$\text{S.E. of } (p - p_1) = \sqrt{\left(\frac{pq}{n_1 + n_2} \times \frac{n_2}{n_1}\right)} \qquad [c.f.\ (14.5b)] \qquad \ldots (*)$$

where $p = \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 \times 0.75 + 500 \times 0.60}{400 + 500} = 0.67 \quad \Rightarrow \quad q = 1 - 0.67 = 0.33$

$$\therefore \quad \text{S.E. of } (p - p_1) = \sqrt{\left(\frac{0.67 \times 0.33}{400 + 500} \times \frac{500}{400}\right)} = 0.018 \text{ [Using (*)]}$$

Test Statistic. Under the null hypothesis, H_0, the test statistic is :

$$Z = \frac{p - p_1}{\text{S.E. of } (p - p_1)} \sim N(0, 1), \qquad \text{(Since samples are large.)}$$

$$\therefore \quad Z = \frac{0.67 - 0.75}{0.018} = \frac{-0.08}{0.018} = 4.08$$

Conclusion. Since the calculated value of $|Z|$ is greater than 3, it is significant. Hence the null hypothesis H_0 is rejected and we conclude that there is significant difference between p_1 and $p = \hat{P}$.

Example 14.16. *If for one-half of n events, the chance of success is p and the chance of failure is q, while for the other half the chance of success is q and the chance of failure is p, show that the standard deviation of the number of successes is the same as if the chance of successes were p in all the cases, i.e., $\sqrt{npq}$ but that the mean of the number of successes is n/2 and not np.*

Solution. Let X_1 and X_2 denote the number of successes in the first half and the second half of n events respectively. Then according to the given conditions, we have

$$\left.\begin{aligned} E(X_1) &= \tfrac{n}{2} p \\ V(X_1) &= \tfrac{n}{2} pq \end{aligned}\right\} \quad \text{and} \quad \left.\begin{aligned} E(X_2) &= \tfrac{n}{2} q \\ V(X_2) &= \tfrac{n}{2} qp \end{aligned}\right\}$$

The mean and variance of the number of successes in all the n events is given by :

$$E(X_1 + X_2) = E(X_1) + E(X_2) = \tfrac{n}{2} p + \tfrac{n}{2} q = \tfrac{n}{2}$$

and $$V(X_1 + X_2) = V(X_1) + V(X_2) = \tfrac{n}{2} pq + \tfrac{n}{2} qp = npq.$$

since the first and second half of events are independent.

Hence the variance is the same as if the probability of success in all the n events is p.

14.8. SAMPLING OF VARIABLES

In the present section we will discuss in detail the sampling of variables, *i.e.*, height, weight, age, income, etc. In the case of sampling of variables each unit of the population provides the value of the variable and the aggregate of these values forms the frequency distribution of the population. From the population, a sample of size n can be drawn by any of the sampling methods discussed earlier which is same as choosing n values of the given variable from the distribution.

14.8.1. Unbiased Estimate for Population Mean (μ) and Variance

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a large population $X_1, X_2, \ldots, X_N$ (N) with mean μ and variance σ^2. Then the sample mean ($\bar{x}$) and variance (s^2) are given by :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \text{and} \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Now $$E(\bar{x}) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(x_i)$$

Since x_i is a sample observation from the population X_i, $(i = 1, 2, \ldots, N)$, any one of the values $X_1, X_2, \ldots, X_N$ each with equal probability $1/N$.

$$\therefore \quad E(x_i) = \frac{1}{N} X_1 + \frac{1}{N} X_2 + \ldots + \frac{1}{N} X_N = \frac{1}{N} (X_1 + X_2 + \ldots + X_N)$$

Hence $$E(\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} (\mu) = \frac{1}{n} n\mu \quad \Rightarrow \quad E(\bar{x}) = \mu$$

Thus, *the sample mean* ($\bar{x}$) *is an unbiased estimate of the population mean* (μ).

Now $$E(s^2) = E\left\{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right\} = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = \frac{1}{n} \sum_{i=1}^{n} E(x_i^2) -$$

We have $V(x_i) = E[x_i - E(x_i)]^2 = E(x_i - \mu)^2$,

$$= \frac{1}{N} \left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_N - \mu)^2\right] = \sigma^2$$

Also $V(x) = E(x^2) - [E(x)]^2 \quad \Rightarrow \quad E(x^2) = V(x) + [E(x)]^2$

In particular $E(x_i^2) = V(x_i) + [E(x_i)]^2 = \sigma^2 +$

Also from (4), we obtain $E(\bar{x}^2) = V(\bar{x}) + [E(\bar{x})]^2$

But $V(\bar{x}) = \frac{\sigma^2}{n}$, where σ^2 is the population variance.

$$\therefore \quad E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2$$

Substituting from (5) and (5a) in (2) we get

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \frac{1}{n}\,n(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right) = \left(1-\frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2 \qquad \ldots(14\cdot7)$$

Since $E(s^2) \neq \sigma^2$, sample variance is not an unbiased estimate of population variance.

From (14·7), we get $\quad \frac{n}{n-1}E(s^2) = \sigma^2 \quad \Rightarrow \quad E\left(\frac{ns^2}{n-1}\right) = \sigma^2$

$$\Rightarrow \quad E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} = \sigma^2, \quad \text{i.e.,} \quad E(S^2) = \sigma^2 \qquad \ldots(14\cdot8)$$

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

$\therefore$ *S^2 is an unbiased estimate of the population variance σ^2.* $\qquad \ldots(14\cdot8a)$

Alter for $E(s^2)$.

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

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

But $\quad \sum_i (x_i-\mu) = \sum_i x_i - n\mu = n\bar{x} - n\mu = n(\bar{x}-\mu)$

$$\therefore \quad s^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2$$

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

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

Remarks 1. Here we see that although sample mean is an unbiased estimate of population sample variance is not an unbiased estimate of population variance. However, an d estimate of σ^2 is given by S^2, defined in equation (14·8*a*).

lays a very important role in sampling theory, particularly in small sampling theory. er σ^2 is not known, its estimate S^2 given by (14·8*a*) is used for practical purposes.

e have $\quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

$$ns^2 = (n-1)S^2 \quad \Rightarrow \quad s^2 = \left(1-\frac{1}{n}\right)S^2$$

for large samples, *i.e.*, for $n \to \infty$, we have $s^2 \to S^2$.

r words, for large samples (*i.e.*, $n \to \infty$), we may take $\hat{\sigma}^2 = s^2$ $\qquad \ldots(14\cdot8b)$

14·8·2. Standard Error of Sample Mean. *The variance of the sample mean is σ^2/n, where σ is the population standard deviation and n is the size of the random sample. The S.E. of mean of a random sample of size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.*

Proof. Let x_i $(i = 1, 2, \ldots, n)$ be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is : $\quad \bar{x} = \frac{1}{n}(x_1+x_2+\ldots+x_n)$

$$\therefore \quad V(\bar{x}) = V\left\{\frac{1}{n}(x_1+x_2+\ldots+x_n)\right\} = \frac{1}{n^2}V(x_1+x_2+\ldots+x_n)$$

$$= \frac{1}{n^2}\left\{V(x_1)+V(x_2)+\ldots+V(x_n)\right\},$$

the covariance terms vanish since the sample observations are independent.

But $\quad V(x_i) = \sigma^2, (i = 1, 2, \ldots, n)$ [From (3) of § 14·8·1]

$$\therefore \quad V(\bar{x}) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \quad \Rightarrow \quad \text{S.E.}(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \qquad \ldots(14\cdot9)$$

14·8·3. Test of Significance for Single Mean. We have proved that if x_i $(i = 1, 2, \ldots, n)$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance σ^2/n, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$. However, this result holds, i.e., $\bar{x} \sim N(\mu, \sigma^2/n)$, even in random sampling from non-normal population provided the sample size n is large [*c.f.* Central Limit Theorem]. Thus for large samples, the *standard normal variate* corresponding to $\bar{x}$ is :

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

Under the *null hypothesis* H_0, that the sample has been drawn from a population with mean μ and variance σ^2, *i.e.*, there is no significant difference between the sample mean ($\bar{x}$) and population mean (μ), the test statistic (for large samples), is :

$$\left[Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\right] \qquad \ldots(14\cdot9a)$$

Remarks 1. If the population s.d. σ is unknown then we use its estimate provided by th sample variance given by [See (14·8*b*)]. $\hat{\sigma}^2 = s^2 \quad \Rightarrow \quad \hat{\sigma} = s$ (for large samples)

2. *Confidence limits for μ.* 95% confidence interval for μ is given by :

$$|Z| \leq 1{\cdot}96, \text{ i.e., } \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96 \Rightarrow \bar{x} - 1{\cdot}96(\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + 1{\cdot}96(\sigma/\sqrt{n}) \quad \ldots(14$$

and $\bar{x} \pm 1{\cdot}96\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence lin for μ are $\bar{x} \pm 2{\cdot}58\sigma/\sqrt{n}$ and 98% confidence limits for μ are $\bar{x} \pm 2{\cdot}33\sigma/\sqrt{n}$.

However, in sampling from a finite population of size N, the corresponding 95% and confidence limits for μ are respectively

$$\left[\bar{x} \pm 1{\cdot}96\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{and} \quad \bar{x} \pm 2{\cdot}58\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right] \qquad \ldots($$

3. The confidence limits for any parameter (P, μ, etc.) are also known as its *fiducial li*

Example 14·17. *A sample of 900 members has a mean 3·4 cms. and s.d. 2·61 cms. Is the sample from a large population of mean 3·25 cms. and s.d. 2·61 cms. ?*

If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

Solution. *Null Hypothesis,* (H_0) : The sample has been drawn from the population with mean $\mu = 3{\cdot}25$ cms. and S.D. $\sigma = 2{\cdot}61$ cms.

Alternative Hypothesis, $H_1 : \mu \neq 3{\cdot}25$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, (Since n is large.)

Here, we are given : $\bar{x} = 3{\cdot}4$ cms., $n = 900$ cms., $\mu = 3{\cdot}25$ cms. and $\sigma = 2{\cdot}61$ cms.

$$\therefore \quad Z = \frac{3{\cdot}40 - 3{\cdot}25}{2{\cdot}61/\sqrt{900}} = \frac{0{\cdot}15 \times 30}{2{\cdot}61} = 1{\cdot}73$$

Since $|Z| < 1{\cdot}96$, we conclude that the data don't provide us any evidence against the null hypothesis (H_0) which may, therefore, be accepted at 5% level of significance.

95% fiducial limits for the population mean μ are :

$$\bar{x} \pm 1{\cdot}96\,(\sigma/\sqrt{n}) = 3{\cdot}40 \pm 1{\cdot}96\,(2{\cdot}61/\sqrt{900}) = 3{\cdot}40 \pm 0{\cdot}1705, \text{ i.e., } 3{\cdot}5705 \text{ and } 3{\cdot}2295$$

98% fiducial limits for μ are given by :

$$\bar{x} \pm 2{\cdot}33\frac{\sigma}{\sqrt{n}} = 3{\cdot}40 \pm 2{\cdot}33 \times \frac{2{\cdot}61}{30} = 3{\cdot}40 \pm 0{\cdot}2027, \text{ i.e., } 3{\cdot}6027 \text{ and } 3{\cdot}1973$$

Remark. 2·33 is the value z_1 of Z from standard normal probability integrals, such that

$$P(|Z| > z_1) = 0{\cdot}98 \implies P(Z > z_1) = 0{\cdot}49$$

Example 14·18. *An insurance agent has claimed that the average age of policy-holders who insure through him is less than the average for all agents, which is 30·5 years.*

A random sample of 100 policy-holders who had insured through him gave the following age distribution :

Age last birthday	:	*16—20*	*21—25*	*26—30*	*31—35*	*36—40*
No. of persons	:	12	22	20	30	16

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5% level of significance. You are given that $Z(1{\cdot}645) = 0{\cdot}95$.

Solution. *Null Hypothesis,* $H_0 : \mu = 30{\cdot}5$ years, *i.e.*, the sample mean $(\bar{x})$ and population mean (μ) do not differ significantly.

Alternative Hypothesis, $H_1 : \mu < 30{\cdot}5$ years (Left-tailed alternative).

CALCULATIONS FOR SAMPLE MEAN AND S.D.

Age last birthday	*No. of persons* (f)	*Mid-point* x	$d = \frac{x - 28}{5}$	fd
16—20	12	18	−2	−24
21—25	22	23	−1	−22
26—30	20	28	0	0
31—35	30	33	1	30
36—40	16	38	2	32
Total	$N = 100$			$\Sigma fd = 16$

$$\bar{x} = 28 + \frac{5 \times 16}{100} = 28{\cdot}8 \text{ years}; \quad s = 5 \times \sqrt{\left\{\frac{164}{100} - \left(\frac{16}{100}\right)^2\right\}}$$

Since the sample is large, $\hat{\sigma} = s = 6{\cdot}35$ years.

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim N(0, 1),$$

Now

$$Z = \frac{28{\cdot}8 - 30{\cdot}5}{\sqrt{(6{\cdot}35)^2/100}} = \frac{-1{\cdot}7}{0{\cdot}635} = -2{\cdot}681.$$

Conclusion. Since computed value of $Z = -2{\cdot}681 < -1{\cdot}645$ or $|Z| > 1{\cdot}96$, it is significant at 5% level of significance. Hence, we reject the null hypothesis H_0 (Accept H_1) at 5% level of significance and conclude that the insurance agent's claim that the average age of policy-holders who insure through him is less than the average for all agents, is valid.

Example 14·19. *As an application of Central Limit Theorem, show that* $P(|\bar{X} - \mu| < E) > 0{\cdot}95$, *then the minimum sample size* $n = [(1{\cdot}96)^2 \sigma^2/E^2]$ *where* μ *and* σ^2 *are the mean and variance respectively of the parent population and* $\bar{X}$ *is the mean of the random sample.*

Solution. By Central Limit Theorem, we know that $\bar{x} \sim N(\mu, \sigma^2/n)$ approximately, *i.e.*, for large n.

$$\therefore \quad Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ asymptotically, i.e., for large } n.$$

From normal probability tables, we have $P(|Z| \leq 1{\cdot}96) = 0{\cdot}95$

$$\implies \quad P\left\{\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96\right\} = 0{\cdot}95 \quad \text{or} \quad P\left[|\bar{x} - \mu| \leq 1{\cdot}96\frac{\sigma}{\sqrt{n}}\right]$$

We are given that $P[|\bar{X} - \mu| < E] > 0{\cdot}95$

From (*) and (**), we have $\quad E > \dfrac{1{\cdot}96\sigma}{\sqrt{n}} \implies n > \dfrac{(1{\cdot}96)^2\sigma^2}{E^2}$

Hence minimum sample size n for estimating μ with 95% confidence is given by $n = 3{\cdot}84\,\sigma^2/E^2$, where E is the permissible error.

Remark. The minimum sample size for estimating μ with confidence coefficient $(1-\alpha)$ is $n = \sigma^2 z_\alpha^2 / E^2$, where z_α is the significant value of Z at level of significance α and E is the permissible error in the estimate.

Similarly, the minimum sample size for estimating population proportion P with confidence coefficient $(1-\alpha)$ is given by $n = PQ\, z_\alpha^2 / E^2$, where z_α is the significant value of Z at level of significance and E is the permissible error in the estimate. If P is unknown, we may

Example 14.20. *The mean muscular endurance score of a random sample of 60 subjects is found to be 145 with a s.d. of 40. Construct a 95% confidence interval for the true mean. Is the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence ?*

Solution. In usual notations, we are given : $n = 60$, $\bar{x} = 145$ and $s = 40$.

The confidence limits for true mean (μ) are :

$\bar{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$ ($\sigma^2 = s^2$, since sample is large)

$$= 145 \pm \frac{1.96 \times 40}{\sqrt{60}} = 145 \pm \frac{78.4}{7.75} = 145 \pm 10.12 = 134.88 \quad \text{and} \quad 155.12$$

Hence 95% confidence interval for μ is (134.88, 155.12). In the notations of Example 14.19, we have

$$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1.96 \times 40}{5}\right)^2 = (15.68)^2 = 245.86 \simeq 246.$$

$[\because z_{0.05} = 1.96,\ \hat{\sigma} = s = 40 \text{ and } |\bar{x} - \mu| < 5 = E]$

Example 14.21. *The standard deviation of a population is 2.70 cms. Find the probability that a random sample of size 66 (i) the sample mean will differ from the population mean by 0.75 cm. or more, and (ii) the sample mean will exceed the population mean by 0.75 cm. or more. Given that the value of the standard normal probability integral from 0 to 2.25 is 0.4877.*

Solution. Here we are given $n = 66$, $\sigma = 2.70$ cms. Since n is large, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \ldots(*)$$

We want

$$P[|\bar{x} - \mu| \geq 0.75] = 1 - P[|\bar{x} - \mu| < 0.75]$$

$$= 1 - P\left(\left|\frac{\sigma Z}{\sqrt{n}}\right| < 0.75\right) \qquad [\text{From } (*)]$$

$$= 1 - P\left(|Z| < 0.75\,\frac{\sqrt{n}}{\sigma}\right) = 1 - 2P\left(0 < Z < 0.75\,\frac{\sqrt{n}}{\sigma}\right)$$

$$= 1 - 2P\left(0 < Z < 0.75\,\frac{\sqrt{66}}{2.70}\right) = 1 - 2P\left(0 < Z < \frac{0.75 \times 8.124}{2.70}\right)$$

$$= 1 - 2P(0 < Z < 2.25) = 1 - 2 \times 0.4877 = 0.0246$$

$$P(\bar{x} - \mu > 0.75) = P(Z > 0.75\sqrt{n}/\sigma) = P(Z > 2.25)$$

$$= 0.5 - P(0 < Z < 2.25) = 0.5 - 0.4877 = 0.0123.$$

Example 14.22. *A normal population has a mean of 0.1 and standard deviation of 2.1. Find the probability that mean of a sample of size 900 will be negative.*

Solution. Here we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 0.1$ and $\sigma = 2.1$ and $n = 900$. Since $X \sim N(\mu, \sigma^2)$, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$. The standard normal variate corresponding to $\bar{x}$ is given by :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0.1}{2.1/30} = \frac{\bar{x} - 0.1}{0.07} \Rightarrow \bar{x} = 0.1 + 0.07Z, \text{ where } Z \sim N(0, 1)$$

The required probability p, that the sample mean is negative is given by :

$$p = P(\bar{x} < 0) = P(0.1 + 0.07Z < 0) = P\left(Z < \frac{-0.10}{0.07}\right)$$

$$= P(Z < -1.43) = P(Z \geq 1.43) = 0.5 - P(0 < Z < 1.43) = 0.5 - 0.4236 = 0.0764.$$

(From Normal Probability Table

Example 14.23. *The guaranteed average life of a certain type of electric light bulbs is 1,000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 per cent of the bulbs do not fall short of the guaranteed average by more than 2.5 per cent. What must be the minimum size of the sample ?*

Solution. Here $\mu = 1{,}000$ hours, $\sigma = 125$ hours.

Since we do not want the sample mean to be less than the guaranteed average mean ($\mu = 1{,}000$) by more than 2.5%, we should have

$$\bar{x} > 1{,}000 - 2.5\% \text{ of } 1{,}000 \Rightarrow \bar{x} > 1{,}000 - 25 = 975$$

Let n be the given sample size. Then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ since sample is large.}$$

We want $$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{975 - 1{,}000}{125/\sqrt{n}} = -\frac{\sqrt{n}}{5} \qquad (\because$$

According to the given condition : $P(Z > -\sqrt{n}/5) = 0.90 \Rightarrow P(0 < Z < \sqrt{n}/5$

$$\Rightarrow \frac{\sqrt{n}}{5} = 1.28 \qquad \text{(From Normal Probabili}$$

$$\therefore \quad n = 25 \times (1.28)^2 = 41 \text{ (approx.)}$$

Example 14.24. *A survey is proposed to be conducted to know the annual income of the old Statistics graduates of Delhi University. How large should the sample be taken to estimate the mean monthly earnings within plus and minus Rs. 10,000 at 95% confidence level ? The standard deviation of the annual earnings of the entire population is Rs. 30,000.*

Solution. We are given : σ = Rs. 30,000.

We want : $P[|\bar{x} - \mu| < 10{,}000] = 0.95$

We know that, in sampling from normal population or for large samples from any population, $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence from Normal Probability Tables, we have

$$P[|Z| \leq 1.96] = 0.95 \Rightarrow P\left[\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \leq 1.96\right] = 0.95$$

$$\Rightarrow \quad P\left[\,|\,\bar{x}-\mu\,|\,<1{\cdot}96\times(\sigma/\sqrt{n})\right]=0{\cdot}95 \qquad \ldots(**)$$

From (*) and (**), we get $\dfrac{1{\cdot}96\times\sigma}{\sqrt{n}}=10{,}000 \Rightarrow \dfrac{1{\cdot}96\times 30{,}000}{\sqrt{n}}=10{,}000$

$$\therefore \quad n=(1{\cdot}96\times 3)^2=(5{\cdot}88)^2=34{\cdot}56\simeq 35$$

Aliter. Using Remark to Example 14·19,

$$n=\left(\frac{z_\alpha\cdot\sigma}{E}\right)^2=\left(\frac{1{\cdot}96\times 30{,}000}{10{,}000}\right)^2=35.$$

14·8·4. Test of Significance for Difference of Means. Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then, since sample sizes are large,

$$\bar{x}_1\sim N(\mu_1,\ \sigma_1^2/n_1) \quad \text{and} \quad \bar{x}_2\sim N(\mu_2,\ \sigma_2^2/n_2)$$

Also $\bar{x}_1-\bar{x}_2$, being the difference of two independent normal variates is also a normal variate. The value of Z (*S.N.V.*) corresponding to $\bar{x}_1-\bar{x}_2$ is given by :

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-E(\bar{x}_1-\bar{x}_2)}{S.E.\,(\bar{x}_1-\bar{x}_2)}\sim N(0,1)$$

Under the null hypothesis, $H_0:\mu_1=\mu_2$, *i.e.*, there is no significant difference between the sample means, we get

$$E(\bar{x}_1-\bar{x}_2)=E(\bar{x}_1)-E(\bar{x}_2)=\mu_1-\mu_2=0;\ V(\bar{x}_1-\bar{x}_2)=V(\bar{x}_1)+V(\bar{x}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2},$$

the covariance term vanishes, since the sample means $\bar{x}_1$ and $\bar{x}_2$ are independent.

Thus under $H_0:\mu_1=\mu_2$, the test statistic becomes (for large samples),

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}}\sim N(0,1) \qquad \ldots(14{\cdot}11)$$

Remarks 1. If $\sigma_1^2=\sigma_2^2=\sigma^2$, *i.e.*, if the samples have been drawn from the populations with common S.D. σ, then under $H_0:\mu_1=\mu_2$,

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{(1/n_1)+(1/n_2)}}\sim N(0,1) \qquad \ldots[14{\cdot}11a]$$

2. If in (14·11*a*), σ is not known, then its estimate based on the sample variances is used. If the sample sizes are not sufficiently large, then an unbiased estimate of σ^2 is given by :

$$\hat{\sigma}^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}, \text{ since}$$

$$E(\hat{\sigma}^2)=\frac{1}{n_1+n_2-2}\left\{(n_1-1)\,E(S_1^2)+(n_2-1)\,E(S_2^2)\right\}=\frac{1}{n_1+n_2-2}\left[(n_1-1)\,\sigma^2+(n_2-1)\,\sigma^2\right]=\sigma^2$$

But since sample sizes are large, $S_1^2\simeq s_1^2$, $S_2^2\simeq s_2^2$, $n_1-1\simeq n_1$, $n_2-1\simeq n_2$. Therefore in practice, for large samples, the following estimate of σ^2 without any serious error is used :

$$\hat{\sigma}^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2} \qquad \ldots[14{\cdot}11b]$$

However, if sample sizes are small, then an exact sample test, *t*-test for difference of means (*c.f.* Chapter 16) is to be used.

3. If $\sigma_1^2\neq\sigma_2^2$ and σ_1 and σ_2 are not known, then they are estimated from the samples. This results in some error, which is practically immaterial, if samples are large. The estimates for large samples are given by : $\hat{\sigma}_1^2=S_1^2\simeq s_1^2$ and $\hat{\sigma}_2^2=S_2^2\simeq s_2^2$. (Since $n_1-1\simeq n_1$, $n_2-1\simeq n_2$)

In this case, (14·11) gives $Z=\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}}\sim N(0,1)$

Example 14·25. *The means of two single large samples of 1,000 and 2,000 members are 67·5 inches and 68·0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2·5 inches ? (Test at 5% level of significance).*

Solution. In usual notations, we are given :

$$n_1=1{,}000,\ n_2=2{,}000,\ \bar{x}_1=67{\cdot}5 \text{ inches},\ \bar{x}_2=68{\cdot}0 \text{ inches}$$

Null hypothesis, $H_0:\mu_1=\mu_2$ and $\sigma=2{\cdot}5$ inches, *i.e.*, the samples have been drawn from the same population of standard deviation 2·5 inches.

Alternative Hypothesis, $H_1:\mu_1\neq\mu_2$ (Two-tailed)

Test Statistic. Under H_0, the test statistic is :

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left\{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}}}\sim N(0,1) \qquad \text{(since samples are from same population)}$$

Now $$Z=\frac{67{\cdot}5-68{\cdot}0}{2{\cdot}5\times\sqrt{\left(\frac{1}{1000}+\frac{1}{2000}\right)}}=\frac{-0{\cdot}5}{2{\cdot}5\times 0{\cdot}0387}=-5{\cdot}1$$

Conclusion. Since $|Z|>3$, the value is highly significant and we reject the hypothesis and conclude that samples are certainly not from the same population with standard deviation 2·5.

Example 14·26. *In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.*

Solution. In the usual notations, we are given that

$$n_1=400, \quad \bar{x}_1=\text{Rs. } 250, \quad s_1=\text{Rs. } 40$$

$$n_2=400, \quad \bar{x}_2=\text{Rs. } 220, \quad s_2=\text{Rs. } 55$$

Null hypothesis, $H_0:\mu_1=\mu_2$, *i.e.*, the average weekly food expenditure in the two populations of shoppers are equal.

Alternative Hypothesis, $H_1:\mu_1\neq\mu_2$. (Two-tailed)

Test Statistic. Since samples are large, under H_0, the test statistic is

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}\sim N(0,1)$$

... and σ_2 the population standard deviations are not known, we can take for large samples $\sigma_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$ and then Z is given by :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{250 - 220}{\sqrt{\left\{\frac{(40)^2}{400} + \frac{(55)^2}{400}\right\}}} = 8{\cdot}82 \text{ (approx.)}$$

... Since $|Z|$ is much greater than 2·58, the null hypothesis ($\mu_1 = \mu_2$) is ... at 1% level of significance and we conclude that the average weekly ... of two populations of shoppers in markets A and B differ significantly.

Example 14·27. *The average hourly wage of a sample of 150 workers in a plant 'A' was ... with a standard deviation of Rs. 1·08. The average hourly wage of a sample of 200 ... in plant 'B' was Rs. 2·87 with a standard deviation of Rs. 1·28. Can an applicant ... assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A' ?*

Solution. Let X_1 and X_2 denote the hourly wages (in Rs.) of workers in plant A ... plant B respectively. Then, in usual notations we are given :

$$\left.\begin{aligned} n_1 &= 150, \quad \bar{x}_1 = 2{\cdot}56, \quad s_1 = 1{\cdot}08 = \hat{\sigma}_1 \\ n_2 &= 200, \quad \bar{x}_2 = 2{\cdot}87, \quad s_2 = 1{\cdot}28 = \hat{\sigma}_2 \end{aligned}\right\} \text{(Since samples are large.)}$$

... Hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.*, there is no significant difference between the ... level of wages of workers in plant A and plant B.

... Hypothesis, $H_1 : \mu_2 > \mu_1$ *i.e.*, $\mu_1 < \mu_2$ (Left-tailed test)

... Statistic. Under H_0, the test statistic (for large samples) is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$Z = \frac{2{\cdot}56 - 2{\cdot}87}{\sqrt{\left\{\frac{(1{\cdot}08)^2}{150} + \frac{(1{\cdot}28)^2}{200}\right\}}} = \frac{-0{\cdot}31}{\sqrt{0{\cdot}016}} = \frac{-0{\cdot}31}{0{\cdot}126} = -2{\cdot}46.$$

... region. For a one-tailed test, the critical value of Z at 5% level of ... is 1·645. The critical region for left-tailed test thus consists of all values of ...

... sion. Since calculated value of Z (–2·46) is less than critical value ..., it is significant at 5% level of significance. Hence the null hypothesis is ... at 5% level of significance and we conclude that the average hourly wages ... plant 'B' are certainly higher than those paid by plant 'A'.

... ple 14·28. *In a certain factory there are two independent processes manufacturing ... item. The average weight in a sample of 250 items produced from one process is found ... ozs. with a standard deviation of 12 ozs. while the corresponding figures in a sample ... items from the other process are 124 and 14. Obtain the standard error of difference ... the two sample means. Is this difference significant ? Also find the 99% confidence ... for the difference in the average weights of items produced by the two processes ...*

Solution. In usual notations, we are given :

$$\left.\begin{aligned} n_1 &= 250, \quad \bar{x}_1 = 120 \text{ oz.}, \quad s_1 = 12 \text{ oz.} = \hat{\sigma}_1 \\ n_2 &= 400, \quad \bar{x}_2 = 124 \text{ oz.}, \quad s_2 = 14 \text{ oz.} = \hat{\sigma}_2 \end{aligned}\right\} \text{(Since samples are large.)}$$

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} \quad \text{(Since samles are large.)}$$

$$= \sqrt{\left(\frac{144}{250} + \frac{196}{400}\right)} = \sqrt{0{\cdot}576 + 0{\cdot}490} = 1{\cdot}034$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.*, the sample means do not differ significantly.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)} = \frac{120 - 124}{1{\cdot}034} \sim N(0, 1)$$

$$\therefore \quad |Z| = \frac{4}{1{\cdot}034} = 3{\cdot}87$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected and we conclude that there is significant difference between the sample means.

99% confidence limits for $|\mu_1 - \mu_2|$, *i.e.*, for the difference in the average weights of items produced by two processes, are :

$|\bar{x}_1 - \bar{x}_2| \pm 2{\cdot}58\ \text{S.E.}(\bar{x}_1 - \bar{x}_2) = 4 \pm 2{\cdot}58 \times 1{\cdot}034 = 4 \pm 2{\cdot}67$ (approx.) = 6·67 and 1·33

$$\therefore \quad 1{\cdot}33 < |\mu_1 - \mu_2| < 6{\cdot}67.$$

Example 14·29. *The mean height of 50 male students who showed above average participation in college athletics was 68·2 inches with a standard deviation of 2·5 inches; while 50 male students who showed no interest in such participation had a mean height of 67·5 inches with a standard deviation of 2·8 inches.*

(i) Test the hypothesis that male students who participate in college athletics are taller than other male students.

(ii) By how much should the sample size of each of the two groups be increased in order that the observed difference of 0·7 inches in the mean heights be significant at the 5% level of significance.

Solution. Let X_1 and X_2 denote the height (in inches) of athletic participants and non-athletic participants respectively. In the usual notations, we are given :

$$n_1 = 50, \ \bar{x}_1 = 68{\cdot}2, \ s_1 = 2{\cdot}5 ; \ n_2 = 50, \ \bar{x}_2 = 67{\cdot}5, \ s_2 = 2{\cdot}8$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$.

Alternative Hypothesis, $H_1 : \mu_1 > \mu_2$ (Right-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore \quad Z = \frac{68{\cdot}2 - 67{\cdot}5}{\sqrt{\left\{\frac{(2{\cdot}5)^2}{50} + \frac{(2{\cdot}8)^2}{50}\right\}}} = \frac{0{\cdot}7}{\sqrt{0{\cdot}282}} = \frac{0{\cdot}7}{0{\cdot}53} = 1{\cdot}32$$

For a right-tailed test, the critical (significant) value of Z at 5% level of significance is 1·645.

(*i*) Since the calculated value of $Z(1{\cdot}32)$ is less than the critical value (1·645), it is not significant at 5% level of significance. Hence the null hypothesis is accepted and we conclude that the college athletes are not taller than other male students.

(*ii*) The difference between the mean heights of two groups, each of size n will be significant at 5% level of significance if $Z \geq 1{\cdot}645$

$$\Rightarrow \frac{68{\cdot}2 - 67{\cdot}5}{\sqrt{\left\{\frac{(2{\cdot}5)^2}{n} + \frac{(2{\cdot}8)^2}{n}\right\}}} \geq 1{\cdot}645 \quad \text{or} \quad \frac{0{\cdot}7}{\sqrt{14{\cdot}09/n}} \geq 1{\cdot}645, \text{ i.e., } \frac{0{\cdot}7}{3{\cdot}754/\sqrt{n}} \geq 1{\cdot}645$$

$$\therefore \quad n \geq \left(\frac{1{\cdot}645 \times 3{\cdot}754}{0{\cdot}7}\right)^2 = (8{\cdot}8219)^2 = 77{\cdot}83 \simeq 78$$

Hence the sample size of each of the two groups should be increased by at least $78 - 50 = 28$, in order that the difference between the mean heights of the two groups is significant.

14·8·5. Test of Significance for the Difference of Standard Deviations. If s_1 and s_2 are the standard deviations of two independent samples, then under null hypothesis, $H_0 : \sigma_1 = \sigma_2$, *i.e.*, sample standard deviations don't differ significantly, the statistic

$$Z = \frac{s_1 - s_2}{S.E.\,(s_1 - s_2)} \sim N(0, 1), \text{ for large samples.}$$

But in case of large samples, the S.E. of the difference of the sample standard deviations is given by : $\quad S.E.\,(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$

$$\therefore \quad Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)}} \sim N(0, 1), \qquad \ldots(14{\cdot}12)$$

σ_1^2 and σ_2^2 are usually unknown and for large samples, we use their estimates given by the corresponding sample variances. Hence the test statistic reduces to

$$Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)}} \sim N(0, 1) \text{ (for large samples)} \qquad \ldots(14{\cdot}13)$$

Example 14·30. *Random samples drawn from two countries gave the following data relating to the heights of adult males :*

	Country A	*Country B*
Mean height (in inches)	67·42	67·25
Standard deviation (in inches)	2·58	2·50
Number in samples	1,000	1,200

(i) Is the difference between the means significant ?

(ii) Is the difference between the standard deviations significant ?

Solution. In usual notations, we are given :

$n_1 = 1{,}000, \quad \bar{x}_1 = 67{\cdot}42$ inches, $\quad s_1 = 2{\cdot}58$ inches,

$n_2 = 1{,}200, \quad \bar{x}_2 = 67{\cdot}25$ inches, $\quad s_2 = 2{\cdot}50$ inches.

As in the previous examples (since sample sizes are large), we can take $\hat{\sigma}_1 = s_1 = 2{\cdot}58, \quad \hat{\sigma}_2 = s_2 = 2{\cdot}50$

(*i*) $H_0 : \mu_1 = \mu_2$, *i.e.*, the sample means do not differ significantly.

$H_1 : \mu_1 \neq \mu_2$ (Two tailed).

Under the *Null hypothesis*, H_0, the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N(0, 1), \text{ since samples are large.}$$

Now $\quad Z = \dfrac{67{\cdot}42 - 67{\cdot}25}{\sqrt{\left\{\frac{(2{\cdot}58)^2}{1000} + \frac{(2{\cdot}50)^2}{1200}\right\}}} = \dfrac{0{\cdot}17}{\sqrt{\left(\frac{6{\cdot}66}{1000} + \frac{6{\cdot}25}{1200}\right)}} = 1{\cdot}56$

Conclusion. Since $|Z| < 1{\cdot}96$, null hypothesis may be accepted at 5% level of significance and we may conclude that thre is no significant difference between the sample means.

(*ii*) Under H_0 : that there is no significant difference between sample standard deviations, $\quad Z = \frac{s_1 - s_2}{S.E.\,(s_1 - s_2)} \sim N(0, 1)$, since samples are large.

Now $\quad S.E.\,(s_1 - s_2) = \sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)} = \sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)}$

if σ_1 and σ_2 are not known and $\hat{\sigma}_1 = s_1, \hat{\sigma}_2 = s_2$.

$$\therefore \quad S.E.\,(s_1 - s_2) = \sqrt{\left\{\frac{(2{\cdot}58)^2}{2 \times 1000} + \frac{(2{\cdot}50)^2}{2 \times 1200}\right\}} = 0{\cdot}07746$$

Hence $\quad Z = \dfrac{2{\cdot}58 - 2{\cdot}50}{0{\cdot}07746} = \dfrac{0{\cdot}08}{0{\cdot}07746} = 1{\cdot}03$

Conclusion. Since $|Z| < 1{\cdot}96$, the data don't provide us any evidence against null hypothesis which may be accepted at 5% level of significance. Hence the two standard deviations do not differ significantly.

Example 14·31. *Two populations have their means equal, but S.D. of one is twice the other. Show that (i) in the samples of size 2,000 from each drawn under simple sampling conditions, the difference of means will, in all probability, not exceed 0·15 of the smaller S.D. (ii) What is the probability that the difference will exceed half this amount ?*

Solution. Let the standard deviations of the two populations be σ and 2σ respectively and let μ be the mean of each of the two populations. Also we are given $n_1 = n_2 = 2{,}000$. If $\bar{x}_1$ and $\bar{x}_2$ be the two sample means then, since samples are large

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.\,(\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Now $\quad E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu - \mu = 0$

and $\quad S.E.\,(\bar{x}_1 - \bar{x}_2) = \sqrt{\left\{\frac{\sigma^2}{n_1} + \frac{(2\sigma)^2}{n_2}\right\}} = \sigma \cdot \sqrt{\left(\frac{1}{2000} + \frac{4}{2000}\right)} = 0{\cdot}05\sigma$

Substituting in (*), we get

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{0{\cdot}05\,\sigma} \sim N(0, 1)$$

Under simple sampling conditions, we should in all probability have

$|Z| < 3 \Rightarrow |\bar{x}_1 - \bar{x}_2| < 3 \times 0{\cdot}05\sigma = 0{\cdot}15\sigma$, as required.

We want $p = P\left[|\bar{x}_1 - \bar{x}_2| > \frac{1}{2} \times 0{\cdot}15\sigma\right]$

$$p = P[0{\cdot}05\sigma < |Z| < 0{\cdot}075\sigma] \qquad \left\{\because Z = \frac{\bar{x}_1 - \bar{x}_2}{0{\cdot}05\sigma} \sim N(0, 1)\right\}$$

$$= P[|Z| > 1{\cdot}5] = 1 - P[|Z| \leq 1{\cdot}5]$$

$$= 1 - 2P(0 \leq Z \leq 1{\cdot}5) = 1 - 2 \times 0{\cdot}4332 = 0{\cdot}1336.$$

DISCUSSION AND REVIEW QUESTIONS

1. (*a*) There are two populations and P_1 and P_2 are the proportions of members in the two populations belonging to 'low-income' group. It is desired to test the hypothesis $H_0 : P_1 = P_2$. Explain clearly, the procedure that you would follow to carry out the above test at 5% level of significance.

(*b*) State the theorem on which the above test is based.

(*c*) In respect of the above two populations, if it is claimed that P_1, the proportion of 'low-income' group in the first population, is greater than P_2, how will you modify the procedure to test this claim (at 5% level) ?

2. (*a*) Take a concrete illustration and in relation to this illustration, explain the following terms :

(*i*) Null hypothesis and alternative hypothesis. (*ii*) Type *I* and Type *II* errors.

(*iii*) Critica region. (*iv*) Level of significance.

(*b*) Explain clearly the terms "Standard Error" and "Sampling Distribution". Show that in a series of n independent trials with constant probability p of success, the standard error of the distribution of successes is $\sqrt{pq/n}$, where $q = 1 - p$.

(*c*) n individuals fall into one or the other two categories with probabilities p and q $(= 1 - p)$, respectively. The number in the two categories are x_1 and x_2 $(x_1 + x_2 = n)$. Show that covariance between x_1 and x_2 is $(-npq)$. Hence obtain the variance of the difference $[(x_1/n) - (x_2/n)]$, between the two proportions.

3. Explain clearly the procedure generally followed in testing of a hypothesis. Point out the difference between one-tail and two-tail tests.

4. What do you mean by interval estimation and how would you set up the confidence limits for a parameter from a sample ? Give the formula for 95% confidence limits for mean and proportion. What modifications do you have to make if the sampling is done from finite population, (*i*) without replacement, and (*ii*) with replacement ?

5. P_1 and P_2 are the (unknown) proportions of students wearing glasses in two universities A and B. To compare P_1 and P_2, samples of sizes n_1 and n_2 are taken from the two populations and number of students wearing glasses is found to be x_1 and x_2 respectively. Suggest an unbiased estimate of $(P_1 - P_2)$ and obtain its sampling distribution when n_1 and n_2 are large. Explain how to test the hypothesis that $P_1 = P_2$.

6. Define standard error. Obtain standard error of mean

(*b*) Find the standard error of a linear function of a number standard error of the mean of n uncorrelated variables following the

5. Derive the expressions for the standard error of

(*i*) the mean of a random sample of size n, and

(*ii*) the difference of the means of two independent random sam

6. What is meant by a statistical hypothesis ? What are the two types of e that arise in testing a hypothesis ? Briefly explain how a statistical hypothesis is teste

7. Obtain the expressions for the standard error of sampling distributions of : (*i*) sa mean $(\bar{x})$, and (*ii*) sample variance (s^2), in random sampling from a large population. Assume that n, the sample size, is large.

8. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population which has a finite fourth moment $\mu_r = E(X_i - \mu)^r$, $r = 4$; $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$; and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Show that : (*i*) $S^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$, (*ii*) $\text{Var}(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right]$.

(*iii*) $\text{Cov}(\bar{X}, S^2) = \frac{\mu_3}{n}$.

ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

14-1. (*a*) A coin is tossed 10,000 times and it turns up head 5,195 items. Discuss whether the coin may be regarded as unbiased one, explaining briefly the theoretical principles you would use for this purpose.

(*b*) A biased coin was thrown 400 times and head resulted 240 times. Find the standard error of the observed proportion of heads and deduce that the probability of getting a head in a single throw of the coin lies almost certainly between 0·53 and 0·67.

(*c*) Experience has shown that 20% of a manufactured product is of the top quality. In one day's production of 400 articles only 50 are of top quality. Show that either the production of the day taken was not a representative sample or the hypothesis of 20% was wrong.

14-2. (*a*) In a large consignment of oranges a random sample of 64 oranges revealed that 14 oranges were bad. Is it reasonable to assume that 20% of the oranges were bad ?

(*b*) By a mobile court checking in certain buses it was found that out of 1,000 people checked on a certain day at Red Fort, 10 persons were found to be ticketless travellers. If daily 1 lakh passengers travel by the buses, find out the estimated limits to the ticketless travellers.

(*c*) In a random sample of 81 items taken from a large consignment some were found to be defective. If the standard error of the proportion of defective items in the sample is 1/18, find 95% confidence limits of the percentage of defective items in the consignment.

14-3. (*a*) In some dice throwing experiments Weldon threw die 75, 145 times and of these 49,152 yielded a 4, 5 or 6. Is this consistent with the hypothesis that the die was unbiased ?

(*b*) 1,000 apples are taken from a large consignment and 100 are found to be bad. Estimate the percentage of bad apples in the consignment and assign the limits within which the percentage lies.

14-4. (*a*) A personnel manager claims that 80 per cent of all single women hired for secretarial job get married and quit work within two years after they are hired. Test this hypothesis at 5% level of significance if among 200 such secretaries, 112 got married within two years after they were hired and quit their jobs.

unit-V

16·5. F-DISTRIBUTION
16·5·1. Derivation of Snedecor's F-distribution
16·5·2 Constants of F-distribution
16·5·3. Mode and Points of inflexion of F-Distribution
16·6. APPLICATIONS OF F-DISTRIBUTION
16·6·1. F-test for Equality of Two Population Variances
16·6·2. F-test for Testing the Significance of an Observed Multiple Correlation Coefficient
16·6·3. F-rest for Testing the Significance of an Observed Sample Correlation Ratio
16·6·4. F-test for Testing the Linearity of Regression
16·6·5. F-test for Equality of Several Means
16·7. RELATION BETWEEN t AND F DISTRIBUTIONS
16·8. RELATION BETWEEN F AND χ^2 DISTRIBUTIONS
16·9. FISHER'S z-DISTRIBUTION
16·10. FISHER'S Z-TRANSFORMATION
16·10·1. Applications of Z-transformation
DISCUSSION & REVIEW QUESTIONS/
ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

16·1. INTRODUCTION

The entire large sample theory was based on the application of *"Normal Test"* (*c.f.* § 14·9). However, if the sample size n is small, the distribution of the various statistics, *e.g.*, $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ or $Z = (X - nP)/\sqrt{nPQ}$ etc., are far from normality and as such *'normal test'* cannot be applied if n is small. In such cases exact sample tests, pioneered by W.S. Gosset (1908) who wrote under the pen name of Student, and later on developed and extended by Prof. R.A. Fisher (1926), are used. In the following sections we shall discuss: (*i*) *t-test*, (*ii*) *F-test*, and (*iii*) *Fisher's z-transformation.*

The exact sample tests can, however, be applied to large samples also though the converse is not true. In all the exact sample tests, the basic assumption is that *"the population(s) from which sample(s) is (are) drawn is (are) normal, i.e., the parent population(s) is (are) normally distributed."*

16·2. STUDENT'S 't' DISTRIBUTION

Let x_i $(i = 1, 2, \ldots, n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by the statistic :

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \qquad \ldots(16\cdot1)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $\ldots(16\cdot1a)$

is an unbiased estimate of the population variance σ^2, and it follows Student's t-

$$f(t) = \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2},\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}};\ -\infty < t < \infty \qquad \ldots(16\cdot2)$$

Remarks 1. A statistic t following Student's t-distribution with ν *d.f.* will be abbreviated as $t \sim t_\nu$.

2. If we take $\nu = 1$ in (16·2), we get :

$$f(t) = \frac{1}{B\left(\frac{1}{2},\frac{1}{2}\right)}\cdot\frac{1}{(1+t^2)} = \frac{1}{\pi}\cdot\frac{1}{(1+t^2)};\ -\infty < t < \infty \qquad [\because \Gamma(1/2) = \sqrt{\pi}]$$

which is the *p.d.f.* of standard Cauchy distribution. Hence, when $\nu = 1$, Student's t distribution reduces to Cauchy distribution.

16·2·1. Derivation of Student's t-distribution. The expression (16·1) can be re-written as :

$$t^2 = \frac{n(\bar{x}-\mu)^2}{S^2} = \frac{n(\bar{x}-\mu)^2}{ns^2/(n-1)} \Rightarrow \frac{t^2}{(n-1)} = \frac{(\bar{x}-\mu)^2}{\sigma^2/n}\cdot\frac{1}{ns^2/\sigma^2} = \frac{(\bar{x}-\mu)^2/(\sigma^2/n)}{ns^2/\sigma^2}$$

Since x_i $(i = 1, 2, \ldots, n)$ is a random sample from the normal population with mean μ and variance σ^2, $\bar{x} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$

Hence $\frac{(\bar{x}-\mu)^2}{\sigma^2/n}$, being the square of a standard normal variate is a chi-square variate with 1 *d.f.*

Also $\frac{ns^2}{\sigma^2}$ is a χ^2-variate with $(n-1)$ *d.f.* (*c.f. Theorem* 15·5).

Further since $\bar{x}$ and s^2 are independently distributed (*c.f. Theorem* 15·5), $\frac{t^2}{n-1}$ being the ratio of two independent χ^2-variates with 1 and $(n-1)$ *d.f.* respectively, is $\beta_2\left(\frac{1}{2},\frac{n-1}{2}\right)$ variate and its distribution is given by :

$$dF(t) = \frac{1}{B\left(\frac{1}{2},\frac{\nu}{2}\right)}\cdot\frac{(t^2/\nu)^{\frac{1}{2}-1}}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}}d(t^2/\nu),\ 0 \le t^2 < \infty \qquad [\text{where } \nu = (n-1)$$

$$= \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2},\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}}dt;\ -\infty < t < \infty$$

the factor 2 disappearing since the integral from $-\infty$ to ∞ must be unity. This required probability density function as given in (16·2) of Student's t-distributio $\nu = (n-1)$ *d.f.*

Remarks on Student's 't'. 1. ***Importance of Student's t-distribution in Statist*** Gosset, who wrote under pseudonym (pen-name) of Student defined his t in a slightly way, *viz.*, $t = (\bar{x}-\mu)/s$ and investigated its sampling distribution, somewhat empiric paper entitled *'The Probable Error of the Mean'*, published in 1908. Prof. R.A. Fishe defined his own 't' and gave a rigorous proof for its sampling distribution in 1926. " feature of 't' is that both the statistic and its sampling distribution are functionally in of σ, the population standard deviation.

The discovery of 't' is regarded as a landmark in the history of statistical inference. Before Student gave his 't', it was customary to replace σ^2 in $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, by its unbiased estimate S^2 to give $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$ and then normal test was applied even for small samples. It has been found that although the distribution of t is asymptotically normal for large n (*c.f.* § 16·2·5), it is far from normality for small samples. The Student's t ushered in an era of exact sample distributions (and tests) and since its discovery many important contributions have been made towards the development and extension of small (exact) sample theory.

2. ***Confidence or Fiducial Limits for*** μ. If $t_{0\cdot05}$ is the tabulated value of t for $\nu = (n-1)$ *d.f.* at 5% level of significance, *i.e.*, $P(|t| > t_{0\cdot05}) = 0\cdot05 \Rightarrow P(|t| \le t_{0\cdot05}) = 0\cdot95$,

the 95% confidence limits for μ are given by :

$$|t| \le t_{0\cdot05}, \text{ i.e., } \left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| \le t_{0\cdot05} \Rightarrow \bar{x} - t_{0\cdot05}\cdot\frac{S}{\sqrt{n}} \le \mu \le \bar{x} + t_{0\cdot05}\frac{S}{\sqrt{n}}$$

Thus, 95% confidence limits for μ are : $\bar{x} \pm t_{0\cdot05}\cdot(S/\sqrt{n})$ ...[16·2*a*]

Similarly, 99% confidence limits for μ are : $\bar{x} \pm t_{0\cdot01}\,(S/\sqrt{n})$ ...[16·2*b*]

where $t_{0\cdot01}$ is the tabulated value of t for $\nu = (n-1)$ *d.f.* at 1% level of significance.

16·2·2. Fisher's 't' (*Definition*). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If ξ is a $N(0, 1)$ and χ^2 is an independent chi-square variate with n *d.f.*, then Fisher's t is given by : $$t = \frac{\xi}{\sqrt{\chi^2/n}} \qquad ...(16\cdot3)$$

and it follows Student's 't' distribution with n degrees of freedom.

16·2·3. Distribution of Fisher's 't'. Since ξ and χ^2 are independent, their joint probability differential is given by :

$$dF(\xi, \chi^2) = \frac{1}{\sqrt{2\pi}}\exp\{-\xi^2/2\}\,\frac{\exp(-\chi^2/2)\,(\chi^2)^{\frac{n}{2}-1}}{2^{n/2}\,\Gamma(n/2)}\,d\xi\,d\chi^2$$

Let us transform to new variates t and u by the substitution :

$$t = \frac{\xi}{\sqrt{\chi^2/n}} \quad \text{and} \quad u = \chi^2 \Rightarrow \xi = t\sqrt{u/n} \text{ and } \chi^2 = u$$

Jacobian of transformation J is given by :

$$J = \frac{\partial(\xi, \chi^2)}{\partial(t, u)} = \begin{vmatrix} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{n}}$$

The joint *p.d.f* $g(t, u)$ of t and u becomes :

$$g(t, u) = \frac{1}{\sqrt{2\pi}\;2^{n/2}\,\Gamma(n/2)\sqrt{n}}\exp\left\{-\frac{u}{2}\left(1+\frac{t^2}{n}\right)\right\}u^{\frac{n}{2}-\frac{1}{2}}\,du$$

Since $\psi^2 \ge 0$ and $-\infty < \xi < \infty$, $u \ge 0$ and $-\infty < t < \infty$,

Integrating *w.r. to* 'u' over the range 0 to ∞, the marginal *p.d.f.* $g_1(.)$ of t becomes :

$$g_1(t) = \frac{1}{\sqrt{2\pi}\;2^{n/2}\,\Gamma(n/2)\sqrt{n}}\left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{t^2}{n}\right)\right\}u^{(n-1)/2}\,du\right]$$

$$= \frac{1}{\sqrt{2\pi}\;2^{n/2}\,\Gamma(n/2)\sqrt{n}}\cdot\frac{\Gamma\{(n+1)/2\}}{\left[\frac{1}{2}\left(1+\frac{t^2}{n}\right)\right]^{(n+1)/2}}$$

$$= \frac{\Gamma(n+1)/2]}{\sqrt{n}\,\Gamma(n/2)\Gamma(1/2)}\cdot\frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}},\quad -\infty < t < \infty$$

$$= \frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{(n+1)/2}},\quad -\infty < t < \infty$$

which is the probability density function of Student's t-distribution with n *d.f.*

Remarks 1. In Fisher's 't' the *d.f.* is the same as the *d.f.* of chi-square variate.

2. Student's 't' may be regarded as a particular case of Fisher's 't' as explained below :

Since $\bar{x} \sim N(\mu, \sigma^2/n)$, $\xi = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ ...(*) and $\chi^2 = \frac{ns^2}{\sigma^2} = \sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma^2$ ...(**)

is independently distributed as chi-square variate with $(n-1)$ *d.f.* Hence Fisher's t is given by

$$t = \frac{\xi}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x}-\mu)}{\sigma}\cdot\frac{\sigma}{\sqrt{\Sigma(x_i-\bar{x})^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x}-\mu)}{S} = \frac{\bar{x}-\mu}{S/\sqrt{n}}$$

and it follows Student's t-distribution with $(n-1)$ *d.f.* (*c.f.* Remark 1 above.)

Now, (***) is same as Student's 't' defined in (16·1). Hence Student's 't' is a particular case of Fisher's 't'.

16·2·4. Constants of t-distribution. Since $f(t)$ is symmetrical about origin, *i.e.*, $t = 0$, all the moments of odd order about origin vanish, *i.e.*,

$$\mu'_{2r+1}\text{ (about origin)} = 0\,;\ r = 0, 1, 2, \ldots$$

In particular, μ_1' (about origin) = 0 = Mean

Hence central moments coincide with moments about origin.

$$\therefore \qquad \mu_{2r+1} = 0,\ (r = 1, 2, \ldots)$$

The moments of even order are given by :

$$\mu_{2r} = \mu'_{2r}\text{ (about origin)} = \int_{-\infty}^{\infty} t^{2r} f(t)\,dt = 2\int_0^\infty t^{2r} f(t)\,dt$$

$$= 2\cdot\frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}}\int_0^\infty \frac{t^{2r}}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}\,dt$$

This integral is absolutely convergent if $2r < n$.

Put $1 + \frac{t^2}{n} = \frac{1}{y} \Rightarrow t^2 = \frac{n(1-y)}{y} \Rightarrow 2t\,dt = -\frac{n}{y^2}\,dy$

When $t = 0$, $y = 1$ and when $t = \infty$, $y = 0$. Therefore,

$$\mu_{2r} = \frac{2}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)}\int_1^0 \frac{t^{2r}}{(1/y)^{(n+1)/2}}\cdot\frac{-n}{2ty^2}\,dy$$

$$= \frac{n}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)}\int_0^1 (t^2)^{(2r-1)/2}\,y^{[(n+1)/2]-2}\,dy$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 \left[n\left(\frac{1-y}{y}\right)\right]^{r-\frac{1}{2}} y^{[(n+1)/2]-2}\, dy$$

$$= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 y^{\frac{n}{2}-r-1} (1-y)^{r-\frac{1}{2}}\, dy = \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot B\left(\frac{n}{2}-r, r+\frac{1}{2}\right), n > 2r.$$

$$= n^r \frac{\Gamma[(n/2)-r]\,\Gamma(r+\frac{1}{2})}{\Gamma(1/2)\,\Gamma(n/2)} \qquad \ldots (16\cdot4a)$$

$$= n^r \frac{(r-\frac{1}{2})(r-\frac{3}{2})\ldots\frac{3}{2}\frac{1}{2}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}-r\right)}{\Gamma(1/2)\,[(n/2)-1][(n/2)-2]\ldots[(n/2)-r]\Gamma[(n/2)-r]}$$

$$= n^r \frac{(2r-1)(2r-3)\ldots3\cdot1}{(n-2)(n-4)\ldots(n-2r)}, \frac{n}{2} > r \qquad \ldots (16\cdot4b)$$

In particular

$$\mu_2 = n \cdot \frac{1}{(n-2)} = \frac{n}{n-2}, (n > 2) \qquad \ldots (16\cdot4c)$$

and

$$\mu_4 = n^2 \frac{3\cdot1}{(n-2)(n-4)} = \frac{3n^2}{(n-2)(n-4)}, (n > 4) \qquad \ldots (16\cdot4d)$$

Hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3\left(\frac{n-2}{n-4}\right); (n > 4).$

Remarks 1. As $n \to \infty$, $\beta_1 = 0$ and $\beta_2 = \lim_{n\to\infty} 3\left(\frac{n-2}{n-4}\right) = 3 \lim_{n\to\infty}\left[\frac{1-(2/n)}{1-(4/n)}\right] = 3$ ...(16·4e)

2. Changing r to $(r-1)$ in [14·4(*b*)], dividing and simplifying, we shall get the recurrence relation for the moments as $\frac{\mu_{2r}}{\mu_{2r-2}} = \frac{n(2r-1)}{(n-2r)}, \frac{n}{2} > r$ ... (16·4*f*)

3. ***Moment Generating Function of t-distribution.*** From [16·4(*b*)] we observe that if $t \sim t_n$, then all the moments of order $2r < n$ exist but the moments of order $2r \geq n$ do not exist. Hence the *m.g.f.* of *t*-distribution does not exist.

Example 16·1. *Express the constants y_0, a and m of the distribution :*

$$dF(x) = y_0\left(1 - \frac{x^2}{a^2}\right)^m dx, \quad -a \leq x \leq a \qquad \ldots(*)$$

in terms of its μ_2 and β_2.

Show that if x is related to a variable t by the equation :

$$x = \frac{at}{[2(m+1) + t^2]^{1/2}}, \qquad \ldots(**)$$

then t has Student's distribution with $2(m+1)$ degrees of freedom. Use the transformation to calculate the probability that $t \geq 2$ when the degrees of freedom are 2 and also when 4.

Solution. First of all, we shall determine the constant y_0 from the consideration that total probability is unity.

$$\therefore \quad y_0 \int_{-a}^{a} \left(1 - \frac{x^2}{a^2}\right)^m dx = 1 \quad \Rightarrow \quad 2y_0 \int_0^a \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

(∵ Integrand is an even function of x.)

$$\Rightarrow \quad 2y_0 \int_0^{\pi/2} \cos^{2m}\theta \cdot a\cos\theta\, d\theta = 1, \qquad (x = a\sin\theta)$$

$$\Rightarrow \quad 2ay_0 \int_0^{\pi/2} \cos^{2m+1}\theta\, d\theta = 1$$

But we have the Beta integral, $2\int_0^{\pi/2} \sin^p\theta\cos^q\theta\, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right)$ ...(1)

$$\therefore \quad ay_0 . 2\int_0^{\pi/2} \cos^{2m+1}\theta \sin^0\theta\, d\theta = 1 \quad \Rightarrow \quad ay_0 B(m+1, \tfrac{1}{2}) = 1 \qquad \text{[Using (1)]}$$

$$\Rightarrow \quad y_0 = \frac{1}{a\,B\left(m+1, \frac{1}{2}\right)} \qquad \ldots(2)$$

Since the given probability function is symmetrical about the line $x = 0$, we have as in § 16·2·4, $\mu_{2r+1} = \mu'_{2r+1} = 0$; $r = 0, 1, 2, \ldots$ [∵ Mean = Origin]

The moments of even order are given by :

$$\mu_{2r} = \mu_{2r}' \text{ (about origin)} = \int_{-a}^{a} x^{2r} f(x)\, dx = y_0 \int_{-a}^{a} x^{2r}\left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^a x^{2r}\left(1 - \frac{x^2}{a^2}\right)^m dx = 2y_0 \int_0^{\pi/2} (a\sin\theta)^{2r} \cos^{2m}\theta .\, a\cos\theta\, d\theta, \qquad (x = a\sin\theta)$$

$$= y_0 a^{2r+1} . 2\int_0^{\pi/2} \sin^{2r}\theta . \cos^{2m+1}\theta\, d\theta = y_0 a^{2r+1} B(r+\tfrac{1}{2}, m+1) \qquad \text{[Using (1)]}$$

$$= a^{2r} \frac{B(r+\frac{1}{2}, m+1)}{B(m+1, \frac{1}{2})} = a^{2r} \cdot \frac{\Gamma\left(r+\frac{1}{2}\right)\Gamma\left(m+\frac{3}{2}\right)}{\Gamma\left(m+r+\frac{3}{2}\right)\Gamma\left(\frac{1}{2}\right)} \qquad \ldots(***)$$

In particular, $\mu_2 = a^2 \cdot \frac{\Gamma|m+(3/2)| \cdot \frac{1}{2}\Gamma(1/2)}{|m+(3/2)|\,\Gamma|m+(3/2)|\,\Gamma(1/2)} = \frac{a^2}{2m+3}$

$$\Rightarrow \quad a^2 = (2m+3)\mu_2 \qquad \ldots(3)$$

Also $\mu_4 = a^4 \frac{\Gamma(5/2)}{\Gamma|m+(7/2)|} \times \frac{\Gamma|m+(3/2)|}{\Gamma(1/2)} = \frac{3a^4}{(2m+5)(2m+3)}$ (On simplification)

$$\therefore \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(2m+3)}{(2m+5)} \quad \Rightarrow \quad m = \frac{9-5\beta_2}{2(\beta_2-3)} \qquad \text{(On simplification) ... (}$$

Equations (2), (3) and (4) express the constants y_0, a and m in terms of μ_2 and β_2.

$$x = \frac{at}{[2(m+1)+t^2]^{1/2}} \quad \Rightarrow \quad \frac{x^2}{a^2} = \frac{t^2}{2(m+1)+t^2}$$

i.e., $1 - \frac{x^2}{a^2} = \frac{2(m+1)}{2(m+1)+t^2} = \left(1 + \frac{t^2}{n}\right)^{-1}, \qquad (n = 2m+2)$

Also $dx = a\left[\frac{dt}{(n+t^2)^{1/2}} - t \cdot \frac{1}{2}\frac{2t\,dt}{(n+t^2)^{3/2}}\right] = a\frac{1}{(n+t^2)^{1/2}}\left(1 - \frac{t^2}{n+t^2}\right)dt$

$$= \frac{an}{(n+t^2)^{3/2}}\,dt = \frac{a}{\sqrt{n}} \cdot \frac{1}{[1+(t^2/n)]^{3/2}}\,dt$$

Hence the *p.d.f.* of X transforms to

$$dF(t) = y_0 \frac{1}{\left(1+\frac{t^2}{n}\right)^m} \cdot \frac{a}{\sqrt{n}} \cdot \frac{dt}{\left(1+\frac{t^2}{n}\right)^{3/2}}$$

$$= \frac{1}{a\,B\left(m+1, \frac{1}{2}\right)} \cdot \frac{a}{\sqrt{n}} \cdot \frac{dt}{\left(1+\frac{t^2}{n}\right)^{m+(3/2)}}$$

$$= \frac{1}{\sqrt{n}\,B\left(\frac{n}{2}, \frac{1}{2}\right)} \cdot \frac{dt}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}, -\infty < t < \infty \qquad \ldots(5)$$

which is the probability differential of Student's t-distribution with $n = 2(m+1)$ *d.f.* Hence the result.

For 2 d.f., *i.e.*, $n = 2$, we get $2(m+1) = 2 \Rightarrow m = 0$. Hence from (**), we get (for $m = 0$),

$$x = \frac{at}{(2+t^2)^{1/2}} \quad \Rightarrow \quad x = \frac{\sqrt{2}}{\sqrt{3}}\,a, \text{ when } t = 2.$$

$$\therefore \quad P(t \geq 2) = P\left[X \geq \sqrt{(2/3)}\,a\right] = \int_{a\sqrt{(2/3)}}^{a} dF(x) = \int_{a\sqrt{(2/3)}}^{a} \frac{1}{a\,B(1, \frac{1}{2})}\,dx$$

[From (*), since $m = 0$]

$$= \frac{1}{2a}\left(a - \frac{\sqrt{2}}{\sqrt{3}}a\right) = \frac{\sqrt{3}-\sqrt{2}}{2\sqrt{3}} \qquad \left[\because B\left(1, \tfrac{1}{2}\right) = \frac{\Gamma 1\,\Gamma(1/2)}{\Gamma(3/2)} = \frac{\Gamma(1/2)}{(1/2)\,\Gamma(1/2)} = 2\right]$$

For 4 d.f., *i.e.*, $n = 4$, we get $m = 1$. Proceeding exactly similarly we shall obtain

$$P(t \geq 2) = \frac{1}{2} - \frac{5\sqrt{2}}{16}.$$

Examlpe 16-2. *If the random variables X_1 and X_2 are independent and follow chi-square distribution with n d.f., show that $\sqrt{n}\,(X_1 - X_2)/2\sqrt{X_1X_2}$ is distributed as Student's t with n d.f., independently of $X_1 + X_2$.*

Solution. Since X_1 and X_2 are independent chi-square variates each with n *d.f.*, their joint *p.d.f.* is given by :

$$p(x_1, x_2) = p_1(x_1) \times p_2(x_2)$$

$$= \frac{1}{2^n\,[\Gamma(n/2)]^2} \cdot e^{-(x_1+x_2)/2}\, x_1^{(n/2)-1}\, x_2^{(n/2)-1}; 0 \leq x_1 < \infty, 0 \leq x_2 < \infty$$

Put $\quad u = \dfrac{\sqrt{n}\,(x_1 - x_2)}{2\sqrt{x_1 x_2}} \quad$ and $\quad v = x_1 + x_2$

$$\Rightarrow \quad x_1 = \frac{v}{2}\left[1 + \frac{1}{\sqrt{\left(1+\frac{n}{u^2}\right)}}\right], \quad x_2 = \frac{v}{2}\left[1 - \frac{1}{\sqrt{\left(1+\frac{n}{u^2}\right)}}\right]$$

Jacobian of transformation is : $\quad J = \dfrac{\partial(x_1, x_2)}{\partial(u, v)} = \dfrac{v}{2\sqrt{n}\left(1+\frac{u^2}{n}\right)^{3/2}}$

The joint *p.d.f.* of U and V becomes

$$g(u, v) = p(x_1, x_2)\,|J| = \frac{1}{2^{2n-1}\,\Gamma(n/2)\,\Gamma(n/2)\,\sqrt{n}} \cdot \frac{e^{-v/2}\,v^{n-1}}{\left(1+\frac{u^2}{n}\right)^{(n+1)/2}}; -\infty < u < \infty, 0 \leq v < \infty$$

Using Legender's duplication formula, viz.,

$$\Gamma n = 2^{n-1}\,\Gamma(n/2)\,\Gamma\left(\frac{n+1}{2}\right)/\sqrt{\pi} \quad \Rightarrow \quad \Gamma(n/2) = \frac{\Gamma n\,\sqrt{\pi}}{2^{n-1}\,\Gamma\left(\frac{n+1}{2}\right)}, \text{ we get}$$

$$2^{2n-1}\,\Gamma(n/2)\,\Gamma(n/2)\,\sqrt{n} = \frac{2^{2n-1}\cdot\Gamma n\,\sqrt{\pi}}{2^{n-1}\,\Gamma\left(\frac{n+1}{2}\right)}\,\Gamma\left(\frac{n}{2}\right)\sqrt{n} = 2^n\,\Gamma n\,\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)\left[\because \sqrt{\pi} = \Gamma\frac{1}{2}\right]$$

$$\therefore \quad g(u, v) = \left(\frac{1}{2^n\,\Gamma n}\,e^{-v/2}\,v^{n-1}\right)\left[\frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \frac{1}{\left(1+\frac{u^2}{n}\right)^{(n+1)/2}}\right]; 0 < v < \infty, -\infty < u < \infty$$

$$\Rightarrow \quad g(u, v) = g_1(u)\,g_2(v),$$

where $\quad g_1(u) = \dfrac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \dfrac{1}{\left(1+\frac{u^2}{n}\right)^{(n+1)/2}}, -\infty < u < \infty$

and $\quad g_2(v) = \dfrac{1}{2^n\,\Gamma n}\,e^{-v/2}\,v^{n-1}, 0 < v < \infty$

(*i*) $\Rightarrow U = \sqrt{n}\,(X_1 - X_2)/2\sqrt{X_1X_2}$ and $V = X_1 + X_2$ are independently distributed

(*ii*) $\Rightarrow U = \sqrt{n}\,(X_1 - X_2)/2\sqrt{X_1X_2} \sim t_n$, and

(*iii*) $\Rightarrow V = X_1 + X_2 \sim \gamma(a = \frac{1}{2}, n)$

Example 6-3. *If $I_x(p, q)$ represents the incomplete Beta function defined by :*

$$I_x(p, q) = \frac{1}{B(p, q)}\int_0^x t^{p-1}\,(1-t)^{q-1}\,dt\,;\, p > 0,\, q > 0$$

show that the distribution function $F(.)$ of Student's t-distribution is given by :

$$F(t) = 1 - \frac{1}{2}\,I_x\left(\frac{n}{2}, \frac{1}{2}\right), \text{ where } x = \left(1+\frac{t^2}{n}\right)^{-1}.$$

Solution. If $f(.)$ is *p.d.f.* of Student's t-distribution with n *d.f.*, then

$$F(t) = \int_{-\infty}^{t} f(u)\,du = 1 - \int_{t}^{\infty} f(u)\,du = 1 - \frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)}\int_{t}^{\infty}\left(1+\frac{u^2}{n}\right)^{-(n+1)/2}du$$

Put $\quad 1 + \dfrac{u^2}{n} = \dfrac{1}{z} \quad \Rightarrow \quad u = \sqrt{n}\left(\dfrac{1-z}{z}\right)$

Also $\quad \dfrac{2u\,du}{n} = \dfrac{-dz}{z^2} \quad \Rightarrow \quad du = -\dfrac{n\,dz}{2uz^2} = -\dfrac{\sqrt{n}}{2}\left(\dfrac{z}{1-z}\right)^{1/2}\cdot\dfrac{dz}{z^2}$

Substituting in (**), we get :

$$F(t) = 1 - \frac{1}{\sqrt{n}B\left(\frac{1}{2}, \frac{n}{2}\right)}\int_{1}^{0}\left(1+\frac{t^2}{n}\right)^{-1} z^{(n+1)/2}\left\{-\frac{\sqrt{n}}{2}\,z^{-3/2}\,(1-z)^{-1/2}\right\}dz$$

$$= 1 + \frac{1}{2\,B\left(\frac{1}{2}, \frac{n}{2}\right)}\int_{1}^{0}\left(1+\frac{t^2}{n}\right)^{-1} z^{(n/2)-1}\,(1-z)^{-1/2}\,dz$$

$$= 1 - \frac{1}{2B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^x z^{(n/2)-1} (1-z)^{(1/2)-1} \, dz \qquad \left[\text{where } z = \left(1 + \frac{t^2}{n}\right)^{-1}\right]$$

$$= 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right), \qquad \left[z = \left(1 + \frac{t^2}{n}\right)^{-1}\right]$$

where $I_x(p, q)$ is defined in (*).

Example 16-4. *Show that for t-distribution with n d.f., mean deviation about mean is given by* $\sqrt{n}\ \Gamma[(n-1)/2] \,/\, \sqrt{\pi}\,\Gamma(n/2)$

Solution. $E(t) = 0$.

$$\text{M.D. (about mean)} = \int_{-\infty}^{\infty} |t|\, f(t)\, dt = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{-\infty}^{\infty} \frac{|t|\, dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

$$= \frac{2}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{t\, dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}} = \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{dy}{(1+y)^{(n+1)/2}}, \qquad \left(\frac{t^2}{n} = y\right)$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{y^{1-1}}{(1+y)^{\frac{n-1}{2}+1}} dy = \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot B\left(\frac{n-1}{2}, 1\right) = \frac{\sqrt{n}\,\Gamma[(n-1)/2]}{\sqrt{\pi}\,\Gamma(n/2)}$$

16-2-5. Limiting Form of t-distribution. *As $n \to \infty$, the p.d.f. of t-distribution with n d.f. viz.,*

$$f(t) = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \to \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right), \quad -\infty < t < \infty$$

Proof. $$\lim_{n \to \infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} = \lim_{n \to \infty} \frac{1}{\sqrt{n}} \frac{\Gamma[(n+1)/2]}{\Gamma(1/2)\,\Gamma(n/2)} = \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\pi}} \left(\frac{n}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}$$

$$\left[\because \Gamma(1/2) = \sqrt{\pi} \text{ and } \lim_{n \to \infty} \frac{\Gamma(n+k)}{\Gamma(n)} = n^k, \text{ (c.f. Remark to § 16-8)}\right]$$

$$\therefore \lim_{n \to \infty} f(t) = \lim_{n \to \infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \lim_{n \to \infty} \left[\left(1 + \frac{t^2}{n}\right)^n\right]^{-\frac{1}{2}} \times \lim_{n \to \infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(-t^2/2), \quad -\infty < t < \infty$$

Hence for large d.f. *t*-distribution tends to standard normal distribution.

16-2-6. Graph of t-distribution. The *p.d.f.* of *t*-distribution with *n d.f.* is :

$$f(t) = C\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty$$

Since $f(-t) = f(t)$, the probability curve is symmetrical about the line $t = 0$. As t increases, $f(t)$ decreases rapidly and tends to zero as $t \to \infty$, so that t-axis is an asymptote to the curve. We have shown that

$$\mu_2 = \frac{n}{n-2}, \; n > 2 \; ; \quad \beta_2 = \frac{3(n-2)}{(n-4)}, \; n > 4$$

Hence for $n > 2$, $\mu_2 > 1$ i.e., the variance of *t*-distribution is greater than that of standard normal distribution and for $n > 4$, $\beta_2 > 3$ and thus *t*-distribution is more flat on the top than the normal curve. In fact, for small *n*, we have

$$P(|t| \geq t_0) \geq P(|Z| \geq t_0), \quad Z \sim N(0, 1)$$

i.e., the tails of the *t*-distribution have a greater probability (area) than the tails of standard normal distribution. Moreover we have also seen [§ 16-2-5], that for large *n* (*d.f.*), *t*-distribution tends to standard normal distribution.

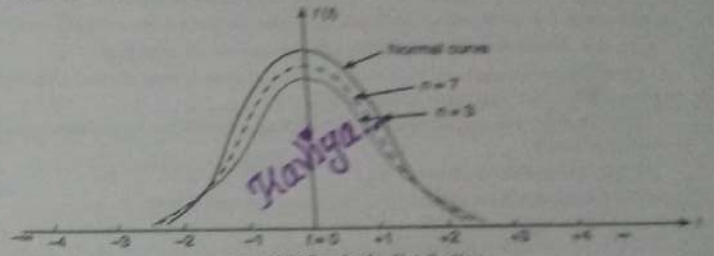


Fig. 16-1 :Graph of t-distribution

16-2-7. Critical Values of t. The critical (or significant) values of *t* at level of significance α and *d.f.* ν for two-tailed test are given by the equation :

$$P[|t| > t_\nu(\alpha)] = \alpha$$

$$\Rightarrow \quad P[|t| \leq t_\nu(\alpha)] = 1 - \alpha$$

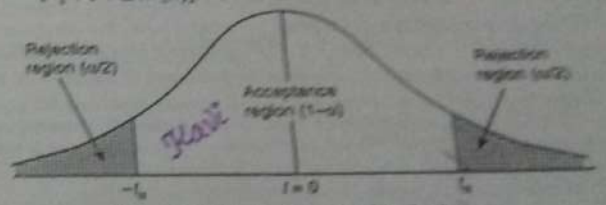


Fig. 16-2 : Critical values of t-distribution

The values $t_\nu(\alpha)$ have been tabulated in Fisher and Yates' Tables, for c values of α and ν and are given in *Table I* at the end of the chapter.

Since *t*-distribution is symmetric about $t = 0$, we get from (16-5)

$$P(t > t_\nu(\alpha)) + P[t < -t_\nu(\alpha)] = \alpha \quad \Rightarrow \quad 2\,P[t > t_\nu(\alpha)] = \alpha$$

$$\Rightarrow \quad P[t > t_\nu(\alpha)] = \alpha/2 \qquad \therefore \; P[t > t_\nu(2\alpha)] = \alpha$$

$t_\nu(2\alpha)$ (from the Tables at the end of the chapter) gives the significant for a single-tail test [Right-tail or Left-tail-since the distribution is symm level of significance α and ν *d.f.*

Hence the significant values of t at level of significance 'α' for a single-taile obtained from those of two-tailed test by looking the values at level of significance

For example,

t_8 (0·05) for single-tail test $= t_8$ (0·10) for two-tail test $= 1{\cdot}86$

t_{15}(0·01) for single-tail test $= t_{15}$ (0·02) for two-tail test $= 2{\cdot}60$.

16·3. APPLICATIONS OF t-DISTRIBUTION

The t-distribution has a wide number of applications in Statistics, some of which are enumerated below.

(*i*) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean :

(*ii*) To test the significance of the difference between two sample means.

(*iii*) To test the significance of an observed sample correlation coefficient and sample regression coefficient.

(*iv*) To test the significance of observed partial correlation coefficient.

In the following sections we will discuss these applications in detail, one by one.

16·3·1. *t*-Test for Single Mean. Suppose we want to test :

(*i*) if a random sample x_i ($i = 1, 2, \ldots, n$) of size n has been drawn from a normal population with a specified mean, say μ_0, or

(*ii*) if the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis, H_0 :

(i) The sample has been drawn from the population with mean μ_0 or

(ii) there is no significant difference between the sample mean $\bar{x}$ and the population mean μ_0,

the statistic
$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}, \qquad \ldots(16\cdot6)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $\qquad \ldots[16\cdot6a)$

follows Student's t-distribution with $(n-1)$ *d.f.*

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t| >$ tabulated t, null hypothesis is rejected and if calculated $|t| <$ tabulated t, H_0 may be accepted at the level of significance adopted.

Remarks* 1. *On computation of S^2 for numerical problems. If $\bar{x}$ comes out in integers, the formula (16·6*a*) can be conveniently used for computing S^2. However, if $\bar{x}$ comes in fractions then the formula (16·6*a*) for computing S^2 is very cumbersome and is not recommended. In that case, step deviation method, given below, is quite useful.

If we take $d_i = x_i - A$, where A is any arbitrary number, then

$$S^2 = \frac{1}{n-1}\left[\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \qquad \ldots(16\cdot6b)$$

$$= \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right], \text{ since variance is independent of change of origin.} \qquad \ldots(16\cdot6c)$$

Also, in this case $\bar{x} = A + \frac{\sum d_i}{n}$. $\qquad \ldots(16\cdot6d)$

2. We know, the sample variance : $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \quad \Rightarrow \quad ns^2 = (n-1)S^2$

$$\therefore \quad \frac{S^2}{n} = \frac{s^2}{n-1} \qquad \ldots(16\cdot6e)$$

Hence for numerical problems, the test statistic (16·6) on using [16·6(*c*)] becomes

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1} \qquad \ldots(16\cdot6f)$$

3. *Assumption for Student's t-test.* The following assumptions are made in the *t*-test :

(*i*) The parent population from which the sample is drawn is normal.

(*ii*) The sample observations are independent, *i.e.*, the sample is random.

(*iii*) The population standard deviation σ is unknown.

***Example* 16·5.** *A machinist is making engine parts with axle diameters of 0·700 inch. A random sample of 10 parts shows a mean diameter of 0·742 inch with a standard deviation of 0·040 inch. Compute the statistic you would use to test whether the work is meeting specifications. Also state how you would proceed further.*

Solution. Here we are given :

$\mu = 0{\cdot}700$ inche, $\bar{x} = 0{\cdot}742$ inche, $s = 0{\cdot}040$ inche and $n = 10$.

Null Hypothesis, $H_0 : \mu = 0{\cdot}700$, *i.e., the product is conforming to specifications.*

Alternative Hypothesis, $H_1 : \mu \neq 0{\cdot}700$

Test Statistic. Under H_0, the test statistic is : $t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{n-1}$

$$\therefore \quad t = \frac{\sqrt{9}\,(0{\cdot}742 - 0{\cdot}700)}{0{\cdot}040} = 3{\cdot}15$$

How to proceed further. Here the test statistic 't' follows Student's t-distribution with $10 - 1 = 9$ *d.f.* We will now compare this calculated value with the tabulated value of t for 9 *d.f.* and at certain level of significance, say 5%. Let this tabulated value be denoted by t_0.

(*i*) If calculated 't', *viz.*, $3{\cdot}15 > t_0$, we say that the value of t is significant which implies that $\bar{x}$ differs significantly from μ and H_0 is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(*ii*) If calculated $t < t_0$, we say that the value of t is not significant, *i.e.*, there is no significant difference between $\bar{x}$ and μ. In other words, the deviation $(\bar{x} - \mu)$ is just due to fluctuations of sampling and null hypothesis H_0 may be retained at 5% level of significance, *i.e.*, we may take the product conforming to specifications.

***Example* 16·6.** *The mean weekly sales of soap bars in departmental stores was 146·3 per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153·7 and showed a standard deviation of 17·2. Was the advertising campaign successful ?*

Solution. We are given : $n = 22$, $\bar{x} = 153{\cdot}7$, $s = 17{\cdot}2$.

Null Hypothesis. The advertising campaign is not successful, *i.e.*, $H_0 : \mu =$

Alternative Hypothesis, $H_1 : \mu > 146{\cdot}3$ (Right-tail).

Test Statistic. Under H_0, the test statistic is : $t = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$

$$\therefore \quad t = \frac{153{\cdot}7 - 146{\cdot}3}{\sqrt{(17{\cdot}2)^2/21}} = \frac{7{\cdot}4 \times \sqrt{21}}{17{\cdot}2} = 9{\cdot}03$$

Conclusion. Tabulated value of t for 21 *d.f.* at 5% level of significance for *one tailed test* is 1·72. Since calculated value is much greater than the tabulated

...ly significant. Hence we reject the null hypothesis and conclude that the ...rtising campaign was definitely successful in promoting sales.

Example 16-7. *A random sample of 10 boys had the following I.Q.'s : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.*

Solution. *Null hypothesis, H_0 : The data are consistent with the assumption of a mean of 100 in the population, i.e., $\mu = 100$.*

Alternative hypothesis, H_1 : $\mu \neq 100$.

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{(\bar{x} - \mu)}{\sqrt{S^2/n}} \sim t_{(n-1)}$,

...ere $\bar{x}$ and S^2 are to be computed from the sample values of I.Q.'s.

TABLE 16-1 : CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
70	- 27·2	739·84
120	22·8	519·84
110	12·8	163·84
101	3·8	14·44
88	- 9·2	84·64
83	- 14·2	201·64
95	- 2·2	4·84
98	0·8	0·64
107	9·8	96·04
100	2·8	7·84
Total 972		1833·60

Here $n = 10$, $\bar{x} = \dfrac{972}{10} = 97{\cdot}2$ and $S^2 = \dfrac{1833{\cdot}60}{9} = 203{\cdot}73$

$$\therefore \quad | t | = \frac{| 97{\cdot}2 - 100 |}{\sqrt{203{\cdot}73/10}} = \frac{2{\cdot}8}{\sqrt{20{\cdot}37}} = \frac{2{\cdot}8}{4{\cdot}514} = 0{\cdot}62$$

Tabulated $t_{0{\cdot}05}$ for $(10 - 1)$, *i.e.*, 9 *d.f.* for two-tailed test is 2·262.

Conclusion. Since calculated t is less than tabulated $t_{0{\cdot}05}$ for 9 *d.f.*, H_0 may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by :

$$\bar{x} \pm t_{0{\cdot}05}\, S/\sqrt{n} = 97{\cdot}2 \pm 2{\cdot}262 \times 4{\cdot}514 = 97{\cdot}2 \pm 10{\cdot}21 = 107{\cdot}41 \text{ and } 86{\cdot}99$$

Hence the required 95% confidence interval is [86·99, 107·41].

Remark. Alter for computing $\bar{x}$ and S^2. Here we see that $\bar{x}$ comes in fractions and as ...ch the computation of $(x - \bar{x})^2$ is quite laborious and time consuming. In this case we use the ...thod of step deviations to compute $\bar{x}$ and S^2, as given below.

x	$d = x - 90$	d^2
70	- 20	400
120	30	900
110	20	400
101	11	121
88	- 2	4
83	- 7	49
95	5	25
98	8	64
107	17	289
100	10	100
Total	$\Sigma d = 72$	$\Sigma d^2 = 2{,}352$

Here $d = x - A$, where $A = 90$. Therefore

$\bar{x} = A + \dfrac{1}{n}\Sigma d = 90 + \dfrac{72}{10} = 97{\cdot}2$ and $S^2 = \dfrac{1}{n-1}\left[\Sigma d^2 - \dfrac{(\Sigma d)^2}{n}\right] = \dfrac{1}{9}\left[2352 - \dfrac{(72)^2}{10}\right] = 203{\cdot}73.$

Example 16-8. *The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches ? Test at 5% significance level assuming that for 9 degrees of freedom P (t > 1·83) = 0·05.*

Solution. *Null Hypothesis, $H_0 : \mu = 64$ inches*

Alternative Hypothesis, $H_1 : \mu > 64$ inches

TABLE 16-2 : CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	70	67	62	68	61	68	70	64	64	66	Total 660
$x - \bar{x}$	4	1	−4	2	−5	2	4	−2	−2	0	0
$(x - \bar{x})^2$	16	1	16	4	25	4	16	4	4	0	90

$$\bar{x} = \frac{\Sigma x}{n} = \frac{660}{10} = 66; \quad S^2 = \frac{1}{n-1}\Sigma (x - \bar{x})^2 = \frac{90}{9} = 10$$

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{66 - 64}{\sqrt{10/10}} = 2,$$

which follows Student's t-distribution with $10 - 1 = 9$ *d.f.*

Tabulated value of t for 9 *d.f.* at 5% level of significance for single (right) tail-test is 1·833. (This is the value $t_{0{\cdot}10}$ for 9 *d.f.* in the two-tailed tables given at the end of the chapter.)

Conclusion. Since calculated value of t is greater than the tabulated value, it is significant. Hence H_0 is rejected at 5% level of significance and we conclude that the average height is greater than 60 inches.

Example 16-9. *A random sample of 16 values from a normal population showed a mean of 41·5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43·5 inches for the population is not reasonable. Obtain 95 per cent and 99 per cent fiducial limits for the same.*

You may use the following information from statistical tables :

$$\nu = 15, \begin{cases} P = 0{\cdot}05, & t = 2{\cdot}131 \\ P = 0{\cdot}01, & t = 2{\cdot}947 \end{cases}$$

Solution. We are given $n = 16$, $\bar{x} = 41{\cdot}5$ inches and $\Sigma(x - \bar{x})^2 = 135$ sq. inches.

$$\therefore \quad S^2 = \frac{1}{n-1}\Sigma(x-\bar{x})^2 = \frac{135}{15} = 9 \quad \Rightarrow \quad S = 3$$

Null Hypothesis, $H_0 : \mu = 43{\cdot}5$ inches, *i.e.*, the data are consistent with the assumption that the mean height in the population is 43·5 inches.

Alternative Hypothesis, $H_1 : \mu \neq 43{\cdot}5$ inches.

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$

$$\therefore \quad |t| = \frac{|41{\cdot}5 - 43{\cdot}5|}{3/4} = \frac{8}{3} = 2{\cdot}667$$

Here number of degrees of freedom is $(16 - 1) = 15$.

We are given : $t_{0{\cdot}05}$ for 15 *d.f.* = 2·131 and $t_{0{\cdot}01}$ for 15 *d.f.* = 2·947.

Conclusion. Since calculated $|t|$ is greater than 2·131, null hypothesis is rejected at 5% level of significance and we conclude that the assumption of mean of 43·5 inches for the population is not reasonable.

Remark. Since calculated $|t|$ is less than 2·947, null hypothesis ($\mu = 43{\cdot}5$) may be accepted at 1% level of significance.

95% fiducial limits for μ : (*d.f.* = 15)

$$\bar{x} \pm t_{0{\cdot}05} \times \frac{S}{\sqrt{n}} = 41{\cdot}5 \pm 2{\cdot}131 \times \frac{3}{4} = 41{\cdot}5 \pm 1{\cdot}598 \quad \Rightarrow \quad 39{\cdot}902 < \mu < 43{\cdot}098$$

99% fiducial limits for μ : (*d.f.* = 15)

$$\bar{x} \pm t_{0{\cdot}01} \times \frac{S}{\sqrt{n}} = 41{\cdot}5 \pm 2{\cdot}947 \times \frac{3}{4} = 43{\cdot}71 \text{ and } 39{\cdot}29 \quad \Rightarrow \quad 39{\cdot}29 < \mu < 43{\cdot}71$$

16·3·2. t-Test for Difference of Means. Suppose we want to test if two independent samples x_i ($i = 1, 2, \ldots, n_1$) and y_j ($j = 1, 2, \ldots, n_2$) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_X and μ_Y respectively.

Under the null hypothesis (H_0) *that the samples have been drawn from the normal populations with means* μ_X *and* μ_Y *and under the assumption that the population variance are equal, i.e.,* $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), the statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_Y)}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \ldots(16{\cdot}7)$$

where $\bar{x} = \dfrac{1}{n_1}\sum_{i=1}^{n_1} x_i, \qquad \bar{y} = \dfrac{1}{n_2}\sum_{j=1}^{n_2} y_j$

and $S^2 = \dfrac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$ $\ldots(16{\cdot}7a)$

is an unbiased estimate of the common population variance σ^2, follows Student's *t*-distribution with $(n_1 + n_2 - 2)$ *d.f.*

Proof. *Distribution of t defined in (16·7).*

$$\xi = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{\sqrt{V(\bar{x} - \bar{y})}} \sim N(0, 1)$$

But $E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_X - \mu_Y$

$V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y}) = \dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2} = \sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$ (By assumption)

[The covariance term vanishes since samples are independent]

$$\therefore \quad \xi = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(*)$$

Let $\chi^2 = \dfrac{1}{\sigma^2}\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{j=1}^{n_2}(y_j - \bar{y})^2\right]$

$$= \left[\sum_i (x_i - \bar{x})^2/\sigma^2\right] + \left[\sum_j (y_j - \bar{y})^2/\sigma^2\right] = \frac{n_1 s_X^2}{\sigma^2} + \frac{n_2 s_Y^2}{\sigma^2} \qquad \ldots(**)$$

Since $n_1 s_X^2/\sigma^2$ and $n_2 s_Y^2/\sigma^2$ are independent χ^2-variates with $(n_1 - 1)$ and $(n_2 - 1)$ *d.f.* respectively, by the additive property of chi-square distribution, χ^2 defined in (**) is a χ^2-variate with $(n_1 - 1) + (n_2 - 1)$, i.e., $n_1 + n_2 - 2$ *d.f.* Further, since sample mean and sample variance are independently distributed, ξ and χ^2 are independent random variables. Hence Fisher's *t* statistic is given by

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{n_1 + n_2 - 2}}}$$

$$= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \times \frac{1}{\left[\frac{1}{n_1 + n_2 - 2}\left\{\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right\}/\sigma^2\right]^{1/2}}$$

$$= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$$

and it follows Student's *t*-distribution with $(n_1 + n_2 - 2)$ *d.f.* (*c.f.* Remark 1, § 16·2·3).

Remarks 1. S^2, defined in (16·7*a*) is an unbiased estimate of the common population variance σ^2, since

$$E(S^2) = \frac{1}{n_1 + n_2 - 2}E\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right] = \frac{1}{n_1 + n_2 - 2}E\left[(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)E(S_X^2) + (n_2 - 1)E(S_Y^2)\right] = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2\right] = \sigma^2$$

2. An important deduction which is of much practical utility is discussed below :

Suppose we want to test if : (*a*) two independent samples x_i ($i = 1, 2, \ldots, n_1$) and y_j ($j = 1, 2, \ldots, n_2$), have been drawn from the populations with same means, or (*b*) the two sample means $\bar{x}$ and $\bar{y}$ differ significantly or not.

Under the null hypothesis, H_0 that (*a*) *samples have been drawn from the populations with same means, i.e.,* $\mu_X = \mu_Y$, *or (b) the sample means* $\bar{x}$ *and* $\bar{y}$ *do not differ significantly,* the statistic :

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad [\because \mu_X = \mu_Y \text{ under } H_0] \qquad \ldots(16\cdot8)$$

where symbols are defined in (16·7*a*), follows Student's *t*-distribution with $(n_1 + n_2 - 2)$ *d.f.*

3. *On the assumption of t-test for difference of means.* Here we make the following three fundamental assumptions :

(*i*) Parent populations, from which the samples have been drawn are normally distributed.

(*ii*) The population variances are equal and unknown, *i.e.,* $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), where σ^2 is unknown.

(*iii*) The two samples are random and independent of each other.

Thus before applying *t*-test for testing the equality of means it is theoretically desirable to test the equality of population variances by applying *F*-test. (*c.f* § 16·6·1) If the variances do not come out to be equal then *t*-test becomes invalid and in that case Behren's '*d*'-test based on fiducial intervals is used. For practical problems, however, the assumptions (*i*) and (*ii*) are taken for granted.

16·3·3. Paired t-test for Difference of Means. Let us now consider the case when (*i*) the sample sizes are equal, *i.e.,* $n_1 = n_2 = n$ (say), and (*ii*) the two samples are not independent but the sample observations are paired together, *i.e.,* the pair of observations (x_i, y_i), $(i = 1, 2, \ldots, n)$ corresponds to the same (*i*th) sample unit. The problem is to test if the sample means differ significantly or not.

For example, suppose we want to test the efficacy of a particular drug, say, for inducing sleep. Let x_i and y_i $(i = 1, 2, \ldots, n)$ be the readings, in hours of sleep, on the *i*th individual, before and after the drug is given respectively. Here instead of applying the difference of the means test discussed in § 16·3·2, we apply the paired *t*-test given below.

Here we consider the increments, $d_i = x_i - y_i$, $(i = 1, 2, \ldots, n)$.

Under the null hypothesis, H_0 that *increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments,* the statistic : $t = \dfrac{\bar{d}}{S/\sqrt{n}}$ $\ldots(16\cdot9)$

where $\bar{d} = \dfrac{1}{n}\sum_{i=1}^{n} d_i$ and $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$ $\ldots(16\cdot9a)$

follows Student's *t*-distribution with $(n - 1)$ *d.f.*

Example 16·10. *Below are given the gain in weights (in kgs.) of pigs fed on two diets A and B.*

Gain in weight

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B : 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution. *Null hypothesis,* $H_0 : \mu_X = \mu_Y$, *i.e., there is no significant difference between the mean increase in weight due to diets A and B.*

Alternative hypothesis, $H_1 : \mu_X \neq \mu_Y$ (two-tailed).

Diet A			Diet B		
x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$
25	−3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	−8	64
34	6	36	10	−20	400
24	−4	16	47	17	289
14	−14	196	31	1	1
32	4	16	40	10	100
24	−4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	−12	144
25	−3	9	21	−9	81
			35	5	25
			29	−1	1
			22	−8	64
$\Sigma x = 336$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 380$	$\Sigma y = 450$	$\Sigma(y - \bar{y}) = 0$	$\Sigma(y - \bar{y})^2 = 1{,}410$

$\bar{x} = \dfrac{336}{12} = 28$, $\bar{y} = \dfrac{450}{15} = 30$, $S^2 = \dfrac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right] = 71{\cdot}6$

and $n_1 = 12$, $n_2 = 15$

Under null hypothesis (H_0) :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

$$\therefore \quad t = \frac{28 - 30}{\sqrt{71{\cdot}6\left(\frac{1}{12} + \frac{1}{15}\right)}} = \frac{-2}{\sqrt{10{\cdot}74}} = -0{\cdot}609$$

Tabulated $t_{0{\cdot}05}$ for $(12 + 15 - 2) = 25$ *d.f.* is 2·06.

Conclusion. Since calculated $|t|$ is less than tabulated *t*, H_0 may be accepted a 5% level of significance and we may conclude that the two diets do not diff significantly as regards their effect on increase in weight.

Remark. Here $\bar{x}$ and $\bar{y}$ come out to be integral values and hence the direct method computing $\Sigma(x - \bar{x})^2$ and $\Sigma(y - \bar{y})^2$ is used. In case $\bar{x}$ and (or) $\bar{y}$ comes out to be fractional, th the step deviation method is recommended for computation of $\Sigma(x - \bar{x})^2$ and $\Sigma(y - \bar{y})^2$.

Example 16·11. *Samples of two types of electric light bulbs were tested for length o and following data were obtained :*

	Type I	Type II
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1{,}234$ *hrs.*	$\bar{x}_2 = 1{,}036$ *hrs.*
Sample S.D.'s	$s_1 = 36$ *hrs.*	$s_2 = 40$ *hrs.*

is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life ?

Solution. *Null Hypothesis*, $H_0 : \mu_X = \mu_Y$, *i.e.*, the two types I and II of electric bulbs are indentical.

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$, *i.e.*, type I is superior to type II.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{13},$$

where $S^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right]$

$$= \frac{1}{n_1+n_2-2}\left(n_1\sigma_1^2 + n_2\sigma_2^2\right) = \frac{1}{13}\left[8 \times (36)^2 + 7 \times (40)^2\right] = 1,659{\cdot}08$$

$$\therefore \quad t = \frac{1234 - 1036}{\sqrt{1659{\cdot}08\left(\frac{1}{8} + \frac{1}{7}\right)}} = \frac{198}{\sqrt{1659{\cdot}08 \times 0{\cdot}2679}} = 9{\cdot}39$$

Tabulated value of t for 13 *d.f.* at 5% level of significance for right (single)-tailed test is 1·77. [This is the value of $t_{0\cdot10}$ for 13 *d.f.* from two-tail tables given at the end of the chapter.]

Conclusion. Since calculated 't' is much greater than tabulated 't', it is highly significant and H_0 is rejected. Hence the two types of electric bulbs differ significantly. Further, since $\bar{x}_1$ is much greater than $\bar{x}_2$, we conclude that type I is definitely superior to type II.

Example 16-12. *The heights of six randomly chosen sailors are (in inches) : 63, 65, 68, 69, 71, and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss, the light that these data throw on the suggestion that sailors are on the average taller than soldiers.*

Solution. *If the heights of sailors and soldiers be represented by the variables X and Y respectively then the Null Hypothesis is,* $H_0 : \mu_X = \mu_Y$, *i.e., the sailors are not on the average taller than the soldiers.*

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$ (Right-tailed).

Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{14}$

Sailors			Soldiers		
X	$d = X - A = X - 68$	d^2	Y	$D = Y - B = Y - 66$	D^2
63	−5	25	61	−5	25
65	−3	9	62	−4	16
68	0	0	65	−1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
			70	4	16
			71	5	25
Total	0	60	72	6	36
			73	7	49
			Total	18	186

$$\therefore \quad \bar{x} = A + \frac{\Sigma d}{n_1} = 68 + 0 = 68$$

and $\Sigma(x - \bar{x})^2 = \Sigma d^2 - \frac{(\Sigma d)^2}{n_1}$

$= 60 - 0 = 60$

$\bar{y} = B + \frac{\Sigma D}{n_2} = 66 + \frac{18}{10} = 67{\cdot}8$

and $\Sigma(y - \bar{y})^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{n_2}$

$= 186 - \frac{324}{10} = 153{\cdot}6$

$$S^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right] = \frac{1}{14}(60 + 153{\cdot}6) = 15{\cdot}2571$$

$$\therefore \quad t = \frac{68 - 67{\cdot}8}{\sqrt{15{\cdot}2571\left(\frac{1}{6} + \frac{1}{10}\right)}^{1/2}} = \frac{0{\cdot}2}{\sqrt{15{\cdot}2571 \times 0{\cdot}2667}} = 0{\cdot}099$$

Tabulated $t_{0\cdot05}$ for 14 *d.f.* for single-tail test is 1·76.

Conclusion. Since calculated t is much less than 1·76, it is not at all significant at 5% levels of significance. Hence null hypothesis may be retained at 5% level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers.

Example 16-13. *To test the claim that the resistance of electric wire can be reduced by at least 0·05 ohm by alloying, 25 values obtained for each alloyed wire and standard wire produced the following results :*

	Mean	Standard deviation
Alloyed wire	0·083 ohm	0·003 ohm
Standard wire	0·136 ohm	0·002 ohm

Test at 5% level whether or not the claim is substantiated.

Solution. *Null Hypothesis* $H_0 : \mu_1 - \mu_2 \geq 0{\cdot}05$, [*i.e.*, the claim is substantiated]

Alternative Hypothesis $H_1 : \mu_1 - \mu_2 < 0{\cdot}05$ (Left-tailed, test)

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $S^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2-2} = \frac{25 \times (0{\cdot}003)^2 + 25 \times (0{\cdot}002)^2}{25+25-2} = \frac{0{\cdot}000225 + 0{\cdot}0001}{48} = 0{\cdot}0000$

$$\therefore \quad t = \frac{(0{\cdot}083 - 0{\cdot}136) - 0{\cdot}05}{\sqrt{\left[0{\cdot}0000067\left(\frac{1}{25} + \frac{1}{25}\right)\right]}} = -\frac{0{\cdot}103}{0{\cdot}00071} = -145{\cdot}07$$

The (critical) tabulated value of t for 48 *d.f.*, at 5% level of significance tailed test is − 1·645.

Conclusion. Since calculated value of t is much less than tabulated value falls in the rejection region. We, therefore, reject the null hypothesis and conclude the claim is not substantiated.

Example 16-14. *A certain stimulus administered to each of the 12 patients the following increase of blood pressure :*

5, 2, 8, −1, 3, 0, −2, 1, 5, 0, 4 and 6

Can it be concluded that the stimulus will, in general, be accompanied by an blood pressure ?

Solution. Here we are given the increments in blood pressure, *i.e.*, $d_i\ (= x_i - y_i)$.

Null Hypothesis, $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in the blood pressure readings of the patients before and after the drug. In other words, the given increments are just by chance (fluctuations of sampling) and not due to the stimulus.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$, *i.e.*, the stimulus results in an increase in blood pressure.

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$

d	5	2	8	−1	3	0	−2	1	5	0	4	6	
d^2	25	4	64	1	9	0	4	1	25	0	16	36	31 / 185

$\bar{d} = \frac{1}{n}\Sigma d = 2{\cdot}58$ and $S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{11}\left[185 - \frac{(31)^2}{12}\right] = 9{\cdot}5382$

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2{\cdot}58 \times \sqrt{12}}{\sqrt{9{\cdot}5382}} = \frac{2{\cdot}58 \times 3{\cdot}464}{3{\cdot}09} = 2{\cdot}89$$

Tabulated $t_{0{\cdot}05}$ for 11 *d.f.* for single-tail test is 1·80. [This is the value of $t_{0{\cdot}10}$ for 11 in the table for two-tail test given at the end of the chapter.]

Conclusion. Since calculated $t > t_{0{\cdot}05}$, H_0 is rejected at 5% level of significance. Hence we conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

Example 16·15. *In a certain experiment to compare two types of animal foods A and B, the following results of increase in weights were observed in animals :*

Animal number		1	2	3	4	5	6	7	8	Total
Increase in weight in lb	Food A	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

(*i*) *Assuming that the two samples of animals are independent, can we conclude that food B is better than food A ?*

(*ii*) *Also examine the case when the same set of eight animals were used in both the foods.*

Solution. *Null Hypothesis*, H_0 : If the increase in weights due to foods A and B are denoted by X and Y respectively, then $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in increase in weights due to diets A and B.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$ (Left-tailed).

(*i*) If the two samples of animals be assumed to be independent, then we will apply *t*-test for difference of means to test H_0.

Test Statistic. Under $H_0 : \mu_X = \mu_Y$, the test criterion is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

	Food A			Food B	
X	$d = X - 50$	d^2	Y	$D = Y - 52$	D^2
49	−1	1	52	0	0
53	3	9	55	3	9
51	1	1	52	0	0
52	2	4	53	1	1
47	−3	9	50	−2	4
50	0	0	54	2	4
52	2	4	54	2	4
53	3	9	53	1	1
Total	7	37		7	23

$\therefore\ \bar{x} = 50 + \frac{7}{8} = 50{\cdot}875$ $\qquad$ $\bar{y} = 52 + \frac{7}{8} = 52{\cdot}875$

and $\Sigma(x-\bar{x})^2 = \Sigma d^2 - \frac{(\Sigma d)^2}{n_1} = 37 - \frac{49}{8} = 30{\cdot}875$

$\Sigma(y-\bar{y})^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{n_2} = 23 - \frac{49}{8} = 16{\cdot}875$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2\right] = \frac{1}{14}(30{\cdot}875 + 16{\cdot}875) = 3{\cdot}41$$

$$\therefore\quad t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{50{\cdot}875 - 52{\cdot}875}{\sqrt{3{\cdot}41\left(\frac{1}{8} + \frac{1}{8}\right)}} = -2{\cdot}17$$

Tabulated $t_{0{\cdot}05}$ for $(8 + 8 - 2) = 14$ *d.f.* for one-tail test is 1·76.

Conclusion. The critical region for the left-tail test is $t < -1{\cdot}76$. Since calculated t is less than −1·76, H_0 is rejected at 5% level of significance. Hence we conclude that the foods A and B differ significantly as regards their effect on increase in weight. Further, since $\bar{y} > \bar{x}$, food B is superior to food A.

(*ii*) If the same set of animals is used in both the cases, then the readings X and Y are not independent but they are paired together and we apply the paired *t*-test for testing H_0.

Under $H_0 : \mu_X = \mu_Y$, the statistic is : $t = \dfrac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$

X	49	53	51	52	47	50	52	53	Total
Y	52	55	52	53	50	54	54	53	
$d = X - Y$	−3	−2	−1	−1	−3	−4	−2	0	−16
d^2	9	4	1	1	9	16	4	0	44

$\bar{d} = \frac{\Sigma d}{n} = \frac{-16}{8} = -2$ and $S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{7}\left(44 - \frac{256}{8}\right) = 1.714$

$$\therefore \quad |t| = \frac{|\bar{d}|}{\sqrt{S^2/n}} = \frac{2}{\sqrt{1.7143/8}} = \frac{2}{0.4629} = 4.32$$

Tabulated $t_{0.05}$ for $(8-1) = 7$ *d.f.* for one-tail test is 1·90.

Conclusion. Here also the observed value of 't' is significant at 5% level of significance and we conclude that food B is superior to food A.

Example 16·16. *Two laboratories carry out independent estimates of a particular chemicals in a medicine produced by a certain firm. A sample is taken from each batch, halved and the separate halves sent to the two laboratories. The following data is obtained :*

No. of samples	10
Mean value of the difference of estimates	0·6
Sum of the squares of the differences (from their means)	20

Is the difference significant ? (Value of t at 5% level for 9 d.f. is 2·262.)

Solution. Let d stand for the difference between the estimates of the chemical between the two halves of each batch, and $\bar{d}$ the mean value of the difference of estimates. In usual notations, we are given :

$$n = 10, \ \bar{d} = 0.6, \ \Sigma(d - \bar{d})^2 = 20$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.,* the difference is insignificant.

Alternative hypothesis, $H_1 : \mu_1 \neq \mu_2$

Test Statistic. Under H_0, the test statistic is : $t = \frac{\bar{d}}{\sqrt{S^2/n}} \sim t_{10-1}$

where $S^2 = \frac{1}{n-1}\Sigma(d - \bar{d})^2 = \frac{20}{9} = 2.22 \quad \therefore \quad t = \frac{0.6}{\sqrt{2.22/10}} = \frac{0.6}{0.471} = 1.274.$

The tabulated value of t at 5% level for 9 *d.f.*, is 2·262 (given).

Conclusion. Since calculated value of t is less than tabulated value of t, it is not significant. Hence, we may accept the null hypothesis and conclude that the difference is not significant.

16·3·4. t-test for Testing the Significance of an Observed Sample Correlation Coefficient. If r is the observed correlation coefficient in a sample of n pairs of observations from a bivariate normal population, then Prof. Fisher proved that under the null hypothesis, $H_0 : \rho = 0$, *i.e., population correlation coefficient is zero,* the statistic :

$$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)} \qquad \ldots(16\cdot10)$$

follows Student's t-distribution with $(n-2)$ *d.f.* (*c.f. Remark to* § 16·4).

If the value of t comes out to be significant, we reject H_0 at the level of significance adopted and conclude that $\rho \neq 0$, *i.e.,* 'r' is significant of correlation in the population.

If t comes out to be non-significant, then H_0 may be accepted and we conclude that variables may be regarded as uncorrelated in the population.

Example 16·17. *(a) A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0·6. Is this significant of correlation in the population ?*

(b) Find the least value of r in a sample of 18 pairs of observations from a bivariate normal population, significant at 5% level of significance.

Solution. (*a*) We set up the null hypothesis, $H_0 : \rho = 0$, *i.e.,* the observed value of correlation coefficient is not significant of any correlation in the population.

Under H_0 : $t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \sim t_{(n-2)} \quad \therefore \quad t = \frac{0.6\sqrt{27-2}}{\sqrt{(1-0.36)}} = \frac{3}{\sqrt{0.64}} = 3.75$

Tabulated $t_{0.05}$ for $(27-2) = 25$ *d.f.* is 2·06.

Conclusion. Since calculated t is much greater than the tabulated t, it is significant and hence H_0 is discredited at 5% level of significance. Thus we conclude that variables are correlated in the population.

(*b*) Here $n = 18$. From the tables $t_{0.05}$ for $(18-2) = 16$ *d.f.* is 2·12.

Under $H_0 : \rho = 0, \quad t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \sim t_{(n-2)} = t_{16}$

In order that the calculated value of t is significant at 5% level of significance, we should have

$$\left|\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}\right| > t_{0.05} \qquad \Rightarrow \left|\frac{r\sqrt{16}}{\sqrt{(1-r^2)}}\right| > 2.12$$

$$\Rightarrow \ 16r^2 > (2.12)^2(1-r^2) \quad \text{or} \quad 20.493r^2 > 4.493 \quad \text{or} \quad r^2 > \frac{4.493}{20.493} = 0.2192$$

Hence $|r| > 0.4682.$

Example 16·18. *A coefficient of correlation of 0·2 is derived from a random sample of 625 pairs of observations. (i) Is this value of r significant ? (ii) What are the 95% confidence limits to the correlation coefficient in the population ?*

Solution. *Under the null hypothesis* $H_0 : \rho = 0$, *i.e., the value of* $r = 0.2$ *is not significant, the test statistic is :* $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

$$\therefore \quad t = \frac{0.2 \times \sqrt{(625-2)}}{\sqrt{(1-0.04)}} = 5.09$$

Since $d.f. = 625 - 2 = 623$, the significant values of t are same as in the normal distribution, *viz.*, $t_{0.05} = 1.96$ and $t_{0.01} = 2.58$. Since calculated t is much greater than these values; it is highly significant. Hence $H_0 : \rho = 0$ is rejected and we conclude that the sample correlation is significant of correlation in the population.

95% Confidence Limits for ρ *(population correlation coefficient)* are :

$$r \pm 1.96 \text{ S.E. } (r) = r \pm 1.96\,(1-r^2)/\sqrt{n} \qquad \text{[Since } n \text{ large]}$$

$$= 0.2 \pm (1.96 \times 0.96/\sqrt{625}) = 0.2 \pm 0.075 = (0.125, 0.275)$$

99% Confidence Limits for ρ are :

$$0.2 \pm 2.58 \times 0.0384 = 0.2 \pm 0.099 = (0.101, 0.299)$$

16·3·5. t-test for Testing the Significance of an Observed Sample Coefficient. Here the problem is to test if a random sample (x_i, y_i), $(i = 1, 2, \ldots$ been drawn from a bivariate normal population in which regression coefficient on X is β.

The line of regression of Y on X (for the given sample) is :

$$Y - \bar{y} = b(X - \bar{x}), \quad b = \frac{\mu_{11}}{\sigma_x^2} \qquad \ldots(*)$$

The estimate of Y for a given value (say) x_i of X as given by line (*) is :

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x})$$

Under the null hypothesis, H_0 that the *population regression coefficient* is β, Prof. R.A. Fisher proved that the statistic,

$$t = (b - \beta)\left[\frac{(n-2)\sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \hat{y}_i)^2}\right]^{1/2} \qquad \ldots(16\cdot11)$$

follows Student's t-distribution with $(n-2)$ *d.f.*

16·3·6. t-test for Testing the Significance of an Observed Partial Correlation Coefficient. Let $r_{12\cdot34\ldots(k+2)}$ be the partial correlation coefficient of order k, observed in a sample of size n from a multivariate normal population. Prof. R.A. Fisher proved that under the null hypothesis $H_0 : \rho_{12\cdot34\ldots(k+2)} = 0$, *i.e.*, the population correlation coefficient is zero, the sampling distribution of $r_{12\cdot34\ldots(k+2)}$ is same as that of r_{12} under the null hypothesis $\rho_{12} = 0$, with sample size n reduced by k. Hence the test statistic for H_0 becomes :

$$t = \frac{r_{12\cdot34\ldots(k+2)}}{\left[1 - r^2_{12\cdot34\ldots(k+2)}\right]^{1/2}}\sqrt{(n-k-2)} \qquad \ldots(16\cdot11a)$$

which follows Student's t-distribution with $(n-k-2)$ *d.f.*

16·4. DISTRIBUTION OF SAMPLE CORRELATION COEFFICIENT WHEN POPULATION CORRELATION COEFFICIENT, $\rho = 0$ (Sawkin's Method).

The sample correlation coefficient, r in a random sample of size n from an uncorrelated bivariate normal population ($\rho = 0$) has the following probability density function :

$$f(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)}\cdot(1 - r^2)^{(n-4)/2}, \quad -1 \le r \le 1 \qquad \ldots(16\cdot12)$$

Proof. Let (x_i, y_i), $(i = 1, 2, \ldots, n)$ be a random sample of size n drawn from an uncorrelated bivariate normal population ($\rho = 0$) in which $E(X) = E(Y) = 0$ and $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$. Let the variable Y be transformed to the variable Z by means of a linear orthogonal transformation, *viz.*,

$$Z = CY \text{ where } Z_{n\times1} = (z_1, z_2, \ldots, z_n)',$$

$Y_{n\times1} = (y_1, y_2, \ldots, y_n)'$ and $C_{n\times n} = (c_{ij})$, C is an orthogonal matrix.

Let us in particular take $c_{11} = c_{12} = \ldots = c_{1n} = 1/\sqrt{n}$, so that

$$z_1 = \frac{1}{\sqrt{n}}(y_1 + y_2 + \ldots + y_n) = \sqrt{n}\,\bar{y}$$

Now proceeding as in (*Theorem 15·5*), we get $\sum_{i=2}^{n} z_i^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = ns_Y^2 \qquad \ldots(*)$

Since in a bivariate normal distribution, the marginal distributions of X and Y are also normal, we have $Y \sim N(0, \sigma_Y^2)$. Hence by Fisher's Lemma (*Theorem 15·4*) z_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma_Y^2)$.

$$\therefore \quad r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\,s_X\,s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x})}{n\,s_X\,s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{ns_X s_Y}$$

$$\Rightarrow \quad \sqrt{n}\,s_Y r = \frac{\sum(x_i - \bar{x})y_i}{\sqrt{n}\,s_X} = z_2 \text{ (say)}, \qquad \ldots(**)$$

[Since the sum of the squares of coefficients of $y_1, y_2, \ldots, y_n$ in (**) is unity.]

From (*) and (**), we get

$$ns_Y^2 = \sum_{i=2}^{n} z_i^2 = \sum_{i=3}^{n} z_i^2 + z_2^2 = \sum_{i=3}^{n} z_i^2 + nr^2 s_Y^2 \quad \Rightarrow \quad (1 - r^2)n\,s_Y^2 = \sum_{i=3}^{n} z_i^2 \qquad \ldots(***),$$

Since z_i ; $i = 1, 2, \ldots, n$, are independent $N(0, \sigma_Y^2)$, (Z_i/σ_Y) are independent $N(0, 1)$.

Hence from (**), $\quad U = \frac{z_2^2}{\sigma_Y^2} = \frac{nr^2 s_Y^2}{\sigma_Y^2}$

being the square of a standard normal variate is a χ^2-variate with 1 *d.f.* and from (***),

$$V = \sum_{i=3}^{n} z_i^2/\sigma_Y^2 = \sum_{i=3}^{n}(z_i/\sigma_Y)^2 = \frac{(1 - r^2)\,n\,s_Y^2}{\sigma_Y^2},$$

being the sum of squares of $(n-2)$ independent standard normal variates is an independent χ^2-variate with $(n-2)$ *d.f.*

Further, since z_2 and $(z_3, z_4, \ldots, z_n)$ are independent r.v.'s, U and V are independent chi-square variates with 1 and $(n-2)$ *d.f.* respectively.

$$\therefore \quad \frac{U}{U+V} = \frac{nr^2 s_Y^2/\sigma_Y^2}{[nr^2 s_Y^2 + (1 - r^2)\,n\,s_Y^2]/\sigma_Y^2} \sim \beta_1\left(\frac{1}{2}, \frac{n-2}{2}\right) \qquad \text{[cf. Theorem 15·2]}$$

$$\Rightarrow \quad r^2 \sim \beta_1\left(\frac{1}{2}, \frac{n-2}{2}\right)$$

Hence the probability density function of r^2 is given by :

$$dF(r^2) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)}(r^2)^{(1/2)-1}(1 - r^2)^{\frac{n-2}{2}-1}\,d(r^2), \quad 0 \le r^2 \le 1$$

$$\Rightarrow \quad f(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)}(1 - r^2)^{(n-4)/2}, \quad -1 \le r \le 1$$

the factor 2 disappearing from the fact that total probability in the range $-1 \le r \le 1$ must be unity.

Remark. *If $\rho = 0$, then $t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}$ is distributed as Student's t with $(n-2)$ d.f.*

Proof. $\qquad t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$

$$\Rightarrow \quad \frac{t^2}{n-2} = \frac{r^2}{(1-r^2)} = \frac{1}{1-r^2} - 1 \quad \Rightarrow \quad (1 - r^2) = \left[1 + \frac{t^2}{(n-2)}\right]^{-1}$$

From (*i*), we get

$$dt = \sqrt{(n-2)}\,d\left[r/\sqrt{(1-r^2)}\right] \quad \Rightarrow \quad dt = \sqrt{(n-2)}\left[\frac{dr}{\sqrt{(1-r^2)}} + \left(-\frac{1}{2}\right)\frac{(-2r)}{(1-r^2)}\right.$$

$$\Rightarrow \quad dt = \sqrt{(n-2)}\ \frac{dr}{\sqrt{(1-r^2)}}\left(1+\frac{r^2}{1-r^2}\right)$$

$$\Rightarrow \quad dt = \sqrt{(n-2)} \times \frac{dr}{(1-r^2)^{1/2}} \quad \Rightarrow \quad dr = \frac{1}{\sqrt{n-2}}(1-r^2)^{3/2}\,dt$$

As r ranges from -1 to 1, from (i), t ranges from $-\infty$ to ∞.

When $\rho = 0$, the *p.d.f.* of 'r' is given by (16-12) and it transforms to

$$dG(t) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)}(1-r^2)^{\frac{n-4}{2}}\frac{1}{\sqrt{(n-2)}}(1-r^2)^{3/2}\,dt$$

$$= \frac{1}{\sqrt{(n-2)}\,B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{n-2}\right)^{(n-1)/2}} \qquad \text{[From (ii)]}$$

$$= \frac{1}{\sqrt{(n-2)}\,B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{n-2}\right)^{(n-2+1)/2}}, -\infty < t < \infty$$

which is the *p.d.f.* of t-distribution with $(n-2)$ *d.f.*

Hence $\quad t = \frac{r}{\sqrt{(1-r^2)}} \cdot \sqrt{(n-2)} \sim t_{(n-2)}$

Example 16-19. *(a) If (x_i, y_i) is a random sample drawn from an uncorrelated bivariate normal population, derive the distribution of :*

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

(b) Further, when $n = 5$ and if $P(|r| \geq C) = \alpha$, show that C is a root of the equation,

$$C\sqrt{(1-C^2)} + \sin^{-1} C + \frac{\pi(\alpha - 1)}{2} = 0$$

Solution. *(a) c.f.* § 16-4, page 16-26.

(b) $P(|r| \geq C) = 1 - P(|r| \leq C) = 1 - P(-C \leq r \leq C)$

$$= 1 - 2P(0 \leq r \leq C) = 1 - 2\int_0^C f(r)\,dr$$

[$\because f(r)$ is symmetrical about $r = 0$]

When $n = 5$, $\quad f(r) = \frac{1}{B\left(\frac{1}{2}, \frac{3}{2}\right)} \cdot (1-r^2)^{1/2}$ [*c.f.* Equation 16-12, page 16-26]

$$\therefore \quad P(|r| \geq C) = 1 - 2\frac{\Gamma(2)}{\Gamma(1/2)\Gamma(3/2)}\int_0^C (1-r^2)^{1/2}\,dr$$

$$= 1 - 2 \times \frac{1}{\frac{1}{2}\pi}\left[\frac{1}{2}r(1-r^2)^{1/2} + \frac{1}{2}\sin^{-1} r\right]_0^C \qquad \left[\because \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}\right]$$

$$= 1 - \frac{4}{\pi}\left[\frac{1}{2}C(1-C^2)^{1/2} + \frac{1}{2}\sin^{-1} C\right] = \alpha \qquad \text{(Given)}$$

$$\therefore \quad 1 - \frac{2}{\pi}\left[C(1-C^2)^{\frac{1}{2}} + \sin^{-1} C\right] = \alpha \quad \Rightarrow \quad C(1-C^2)^{1/2} + \sin^{-1} C + (\alpha - 1)\frac{\pi}{2} = 0.$$

16-5. F-DISTRIBUTION

Definition. *If X and Y are two independent chi-square variates with ν_1 and ν_2 respectively, then F-statistic is defined by*

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (ν_1, ν_2) *d.f.* with probability function given by :

$$f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, 0 \leq F < \infty$$

Remarks 1. The sampling distribution of F-statistic does not involve any parameters and depends only on the degrees of freedom ν_1 and ν_2.

2. A statistic F following Snedecor's F-distribution with (ν_1, ν_2) *d.f.* will be written as $F \sim F(\nu_1, \nu_2)$.

16-5-1 Derivation of Snedecor's F-distribution. Since X and Y are independent chi-square variates with ν_1 and ν_2 *d.f.* respectively, their joint probability function is given by :

$$f(x, y) = \left\{\frac{1}{2^{\nu_1/2}\Gamma(\nu_1/2)}\exp(-x/2)\,x^{(\nu_1/2)-1}\right\} \times \left\{\frac{1}{2^{\nu_2/2}\Gamma(\nu_2/2)}\exp(-y/2)\,y^{(\nu_2/2)-1}\right\}$$

$$= \frac{1}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\exp\{-(x+y)/2\} \times x^{(\nu_1/2)-1}\,y^{(\nu_2/2)-1}, 0 \leq x, y < \infty$$

Let us make the following transformation of variables :

$F = \frac{x/\nu_1}{y/\nu_2}$ and $u = y$, so that $0 \leq F < \infty, 0 < u < \infty$ $\quad \therefore \quad x = \frac{\nu_1}{\nu_2}Fu$ and $y = u$

Jacobian of transformation J is given by : $\quad J = \frac{\partial(x, y)}{\partial(F, u)} = \begin{vmatrix} \frac{\nu_1}{\nu_2}u & \frac{\nu_1}{\nu_2}F \\ 0 & 1 \end{vmatrix}$

Thus the joint *p.d.f.* of the transformed variables is :

$$g(F, u) = \frac{1}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\} \times \left(\frac{\nu_1}{\nu_2}Fu\right)^{(\nu_1/2)-1} u^{(\nu_2/2)-1}$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\} \times u^{[(\nu_1+\nu_2)/2]-1}\,F^{(\nu_1/2)-1};\ 0 < u < \infty, 0 \leq F < \infty$$

...grating w.r. to u over the range 0 to ∞, the p.d.f. of F becomes :

$$g_1(F) = \frac{(\nu_1/\nu_2)^{(\nu_1/2)}\, F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \times \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\} u^{[(\nu_1+\nu_2)/2]-1}\, du\right]$$

$$= \frac{(\nu_1/\nu_2)^{(\nu_1/2)}\, F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \times \frac{\Gamma[(\nu_1+\nu_2)/2]}{\left[\frac{1}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right]^{(\nu_1+\nu_2)/2}}$$

$$g_1(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \quad 0 \le F < \infty$$

...the required probability function of F-distribution with (ν_1, ν_2) d.f.

...ver,

$$F = \frac{x/\nu_1}{y/\nu_2}$$

$\frac{\nu_1}{\nu_2}F = \frac{x}{y}$, being the ratio of two independent chi-square variates with ... d.f. respectively is a $\beta_2\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$ variate. Hence the probability function of F ...given by :

$$dP(F) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{\left(\frac{\nu_1}{\nu_2}F\right)^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}} d\left(\frac{\nu_1}{\nu_2}F\right)$$

$$\therefore \quad f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \quad 0 \le F < \infty$$

16·5·2. Constants of F-distribution.

$$\mu_r' \text{ (about origin)} = E(F^r) = \int_0^\infty F^r f(F)\, dF$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty F^r \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}} dF \qquad \ldots(*)$$

To evaluate the integral, put : $\frac{\nu_1}{\nu_2}F = y$, so that $dF = \frac{\nu_2}{\nu_1}dy$

$$\mu_r' = \frac{[\nu_1/\nu_2]^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{\left(\frac{\nu_2}{\nu_1}y\right)^{r+(\nu_1/2)-1}}{(1+y)^{(\nu_1+\nu_2)/2}} \left(\frac{\nu_2}{\nu_1}\right) dy$$

$$= \frac{\left(\frac{\nu_2}{\nu_1}\right)^r}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{y^{r+(\nu_1/2)-1}}{(1+y)^{(\nu_1+\nu_2)/2+r+[(\nu_2/2)-r]}} dy$$

$$= \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot B\left(r+\frac{\nu_1}{2}, \frac{\nu_2}{2}-r\right), \quad \nu_2 > 2r \qquad \ldots(16\cdot14)$$

Aliter for (16·14). (16·14) could also be obtained by substituting $\frac{\nu_1}{\nu_2}F = \tan^2\theta$ in (*)

and using the Beta integral : $2\int_0^{\pi/2} \sin^p\theta \cos^q\theta\, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right)$

$$\therefore \quad \mu_r' = \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{\Gamma[r+(\nu_1/2)]\,\Gamma[(\nu_2/2)-r]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}; \; r < \frac{\nu_2}{2} \quad \Rightarrow \quad \nu_2 > 2r \qquad \ldots(16\cdot15)$$

In particular

$$\mu_1' = \frac{\nu_2}{\nu_1} \cdot \frac{\Gamma[1+(\nu_1/2)]\,\Gamma[(\nu_2/2)-1]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2 \qquad \ldots(16\cdot15a)$$

$[\because \Gamma(r) = (r-1)\,\Gamma(r-1)]$

Thus the mean of F-distribution is independent of ν_1.

$$\mu_2' = \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{\Gamma[(\nu_1/2)+2]\,\Gamma[(\nu_2/2)-2]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}$$

$$= \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{[(\nu_1/2)+1]\,(\nu_1/2)}{[(\nu_2/2)-1]\,[(\nu_2/2)-2]} = \frac{\nu_2^2(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)}, \quad \nu_2 > 4.$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \frac{\nu_2^2(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)} - \frac{\nu_2^2}{(\nu_2-2)^2} = \frac{2\nu_2^2(\nu_2+\nu_1-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, \quad \nu_2 > 4 \qquad \ldots(16\cdot15b)$$

Similarly, on putting $r = 3$ and 4 in μ_r', we get μ_3' and μ_4' respectively, from which the central moments μ_3 and μ_4 can be obtained.

Remark. It has been proved that for large degrees of freedom, ν_1 and ν_2, F tends to $N[1, 2\{(1/\nu_1) + (1/\nu_2)\}]$ variate.

16·5·3. Mode and Points of Inflexion of F-distribution. We have

$$\log f(F) = C + \{(\nu_1/2)-1\} \log F - \left(\frac{\nu_1+\nu_2}{2}\right) \log\{1+(\nu_1/\nu_2)F\},$$

where C is a constant independent of F.

$$\therefore \quad \frac{\partial}{\partial F}[\log f(F)] = \left(\frac{\nu_1}{2}-1\right) \cdot \frac{1}{F} - \frac{(\nu_1+\nu_2)}{2} \cdot \frac{1}{\left(1+\frac{\nu_1}{\nu_2}F\right)} \cdot \frac{\nu_1}{\nu_2}$$

$$f'(F) = \frac{\partial}{\partial F} f(F) = 0 \quad \Rightarrow \quad \frac{\nu_1-2}{2F} - \frac{\nu_1(\nu_1+\nu_2)}{2(\nu_2+\nu_1 F)} = 0$$

Hence $$F = \frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)}$$

It can be easily verified that at this point $f''(F) < 0$. Hence mode $= \frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)}$ ...(16·1

Remarks 1. Since $F > 0$, mode exists if and only if $\nu_1 > 2$.

2. Mode $= \left(\frac{\nu_1}{\nu_1+2}\right)\cdot\left(\frac{\nu_1-2}{\nu_1}\right)$

Hence mode of F-distribution is always less than unity.

3. The points of inflexion of F-distribution exist for $\upsilon_1 > 4$ and are equidistant from mode.

Proof. We have $\frac{\nu_1}{\nu_2}F = \frac{X}{Y} \sim \beta_2(l, m)$, ...(*)

where $l = \upsilon_1/2$ and $m = \upsilon_2/2$. We now find the points of inflexion of Beta distribution of second kind with parameters l and m. If $X \sim \beta_2(l, m)$, its *p.d.f.* is :

$$f(x) = \frac{1}{\beta(l, m)}\cdot\frac{x^{l-1}}{(1+x)^{l+m}};\ 0 \le x < \infty$$

Points of inflexion are the solution of $f''(x) = 0$ and $f'''(x) \neq 0$

From (**), $\log f(x) = -\log \beta(l, m) + (l-1)\log x - (l+m)\log(1+x)$

Differentiating twice *w.r. to* x, we get

$$\frac{f'(x)}{f(x)} = \frac{l-1}{x} - \frac{l+m}{1+x}$$

and $$\frac{f(x)f''(x) - [f'(x)]^2}{[f(x)]^2} = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$$

If $f''(x) = 0$, then $-\left[\frac{f'(x)}{f(x)}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2} \Rightarrow -\left[\frac{l-1}{x} - \frac{l+m}{1+x}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$

[On using (***)]

$$\Rightarrow \quad \frac{l-1}{x^2}\left(l-1-1\right) - 2\,\frac{(l-1)(l+m)}{x(1+x)} + \frac{l+m}{(1+x)^2}\times(l+m+1) = 0$$

$$\Rightarrow \quad (l-1)(l-2)(1+x)^2 - 2x(1+x)(l-1)(l+m) + x^2(l+m)(l+m+1) = 0 \qquad ...(****)$$

which is a quadratic in x. It can be easily verified that at these values of x, $f'''(x) \neq 0$, if $l > 2$.

The roots of (****) give the points of inflexion of $\beta_2(l, m)$ distribution. The sum of the points of inflexion is equal to the sum of roots of (****) and is given by :

$$-\left[\frac{\text{Coeff. of } x \text{ in (****)}}{\text{Coeff. of } x^2 \text{ in (****)}}\right] = -\left[\frac{2(l-1)(l-2) - 2(l-1)(l+m)}{(l-1)(l-2) - 2(l-1)(l+m) + (l+m)(l+m+1)}\right]$$

$$= \frac{2(l-1)[(l+m) - (l-2)]}{(l-1)(l-2) - (l-1)(l+m) - (l-1)(l+m) + (l+m)(l+m+1)}$$

$$= \frac{2(l-1)(m+2)}{(l-1)[(l-2-l-m)] + (l+m)(l+m+1-l+1)}$$

$$= \frac{2(l-1)(m+2)}{-(l-1)(m+2) + (l+m)(m+2)} = \frac{2(l-1)}{l+m-l+1} = \frac{2(l-1)}{(m+1)}$$

$\therefore$ Sum of points of inflexion of $\left(\frac{\nu_1}{\nu_2}F\right)$ distribution $= \frac{2(l-1)}{(m+1)} = \frac{2\left(\frac{\nu_1}{2}-1\right)}{\left(\frac{\nu_1}{2}+1\right)} = \frac{2(\nu_1-2)}{(\nu_1+2)}$

$\Rightarrow$ Sum of points of inflexion of $F(\upsilon_1, \upsilon_2)$ distribution

$$= \frac{\nu_2}{\nu_1}\cdot\frac{2(\nu_1-2)}{(\nu_2+2)}, \text{ provided } l = \frac{\nu_1}{2} > 2 = 2\,\frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)} = 2 \text{ Mode, provided } \nu_1 > 4$$

Hence the points of inflexion of $F(\upsilon_1, \upsilon_2)$ distribution, when they exist, (*i.e.*, when $\upsilon_1 > 4$) are equidistant from the mode.

4. Karl Pearson's coefficient of skewness is given by : $\beta_1 = \frac{\text{Mean} - \text{Mode}}{\sigma} > 0$, since mean > 1 and mode < 1. Hence F-distribution is highly positively skewed.

5. The probability $f(F)$ increases steadily at first until it reaches its peak (corresponding to the modal value which is less than 1) and then decreases slowly so as to become tangential to $F = \infty$, *i.e.*, F-axis is an asymptote to the right tail.

Example 16·20. *When $\nu_1 = 2$, show that the significance level of F corresponding to a significant probability p is : $F = \frac{\nu_2}{2}\left[p^{-(2/\nu_2)} - 1\right]$ where υ_1 an υ_2 have their usual meanings.*

Solution. When $\nu_1 = 2$,

$$f(F) = \frac{1}{B\left(1, \frac{\nu_2}{2}\right)}\cdot\frac{2}{\nu_2}\cdot\frac{1}{\left(1+\frac{2}{\nu_2}F\right)^{(\nu_2/2)+1}}$$

$$= \frac{\Gamma\left(\frac{\nu_2}{2}+1\right)}{\Gamma(1)\Gamma(\nu_2/2)}\times\frac{2/\nu_2}{\left(\frac{2}{\nu_2}\right)^{(\nu_2/2)+1}\left(F+\frac{\nu_2}{2}\right)^{(\nu_2/2)+1}} = \frac{\left(\frac{\nu_2}{2}\right)^{(\nu_2/2)+1}}{\left(F+\frac{\nu_2}{2}\right)^{(\nu_2/2)+1}}$$

Hence $$p = \int_F^\infty f(F)\,dF = \left[\frac{\nu_2}{2}\right]^{(\nu_2/2)+1}\times\int_F^\infty\frac{dF}{\left(F+\frac{\nu_2}{2}\right)^{(\nu_2/2)+1}}$$

$$= \left(\frac{\nu_2}{2}\right)^{(\nu_2/2)+1}\times\left|\frac{\left(F+\frac{\nu_2}{2}\right)^{-(\nu_2/2)}}{-\frac{\nu_2}{2}}\right|_F^\infty = \left[\frac{\left(\frac{\nu_2}{2}\right)}{F+\frac{\nu_2}{2}}\right]^{\nu_2/2} = \frac{1}{\left(1+\frac{2}{\nu_2}F\right)^{\nu_2/2}}$$

$$\Rightarrow \quad p^{-(2/\nu_2)} = 1 + \frac{2F}{\nu_2} \quad \Rightarrow \quad F = \frac{\nu_2}{2}\left[p^{-(2/\nu_2)} - 1\right].$$

Example 16·21. *X is a binomial variate with parameters n and p and $F_{\upsilon_1, \upsilon_2}$ is F-statistic with υ_1 and υ_2 d.f. Prove that :*

$$P(X \le k-1) = P\left[F_{2k,\,2(n-k+1)} > \frac{n-k+1}{k}\cdot\frac{p}{1-p}\right].$$

Solution. If $X \sim B(n, p)$, then we have

$$P(X \le k-1) = (n-k+1)\cdot\binom{n}{k-1}\int_0^q t^{n-k}(1-t)^{k-1}\,dt$$

$$= \frac{1}{\beta(k, n-k+1)}\int_0^q t^{n-k}(1-t)^{k-1}\,dt$$

[See Chapter 8—Binomial distribution.]

$$P = P_r\left[F_{2k,2(n-k+1)} > \frac{n-k+1}{k}\left(\frac{p}{1-p}\right)\right] = \int_{\frac{n-k+1}{k}\frac{p}{q}}^\infty p[F_{2k,\,2(n-k+1)}]\,dF$$

$$= \frac{1}{B(k, n-k+1)}\int_{\frac{n-k+1}{k}\frac{p}{q}}^\infty\frac{[k/(n-k+1)]^k\cdot F^{k-1}\,dF}{\left(1+\frac{kF}{n-k+1}\right)^{n+1}}$$