Prepared by Punjab Curriculum & Textbook Board Lahore.
Approved by: Federal Ministry of Education, Curriculum Wing, Islamabad.

CONTENTS

Author:

Prof. Muhammad Rauf Chaudhary
Govt. College for Boys,
Gujranwala.

Editors:

Prof. Muhammad Khalid
Ex-Director (Technical), PCTB

Supervised by:

Madiha Mehmood
Subject specialist, PCTB

Mr. Mazhar Hayat
Subject specialist, PCTB

Artist: Ayesha Waheed

Director (Manuscripts): Mrs. Nisar Qamar

Publisher: Bright Way Publisher, Lahore.
Printer: Qudrat ullah Printers, Lahore.

Date of Printing	Edition	Impression	Copies	Price
Nov.2019	1st	16th	7,000	142.00

10 NORMAL DISTRIBUTION

10.1 NORMAL DISTRIBUTION

The normal distribution is undoubtedly the most important and frequently used of all probability laws, because

(*i*) the normal random variable does frequently occur in practical problems such as heights and weights of individuals, I.Q. scores, errors of measurements, *etc.*

(*ii*) it is the limiting form of many other probability laws and hence provides an accurate approximation to them.

(*iii*) it is also the limiting distribution on the well known *central limit theorem* (as discussed in Theorem 11.6).

Thus a great many techniques used in applied statistics are based on the normal distribution. The formal definition follows.

10.1.1 Normal Probability Density Function. A continuous random variable X is normally distributed if and only if its probability density function is

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

where μ is any real number (*i. e.*, $-\infty < \mu < \infty$) and σ must be positive (*i. e.*, $\sigma > 0$), π ($= 3.141592654\cdots\cdots$) and e ($= 2.718281828\cdots\cdots$) are constants, x is the value of random variable X and $f(x)$ is the density (ordinate) at $X = x$.

A normal distribution is characterized by two parameters μ and σ, its mean and standard deviation respectively. Sometimes it is denoted by $N(\mu, \sigma^2)$. Thus

$$X \sim N(\mu, \sigma^2)$$

means that a random variable X is normally distributed with its mean μ and variance σ^2.

10.1.2 Shape of Normal Distribution. The graph of a normal probability density function is called a normal curve. As can be seen from the definition, the probability density function for a normal random variable has a unimodal, symmetrical and bell shaped distribution.

Figure 10.2 shows the graphs of some typical normal density functions for various values of the parameters μ and σ. The parameter σ controls the relative flatness of the curve.

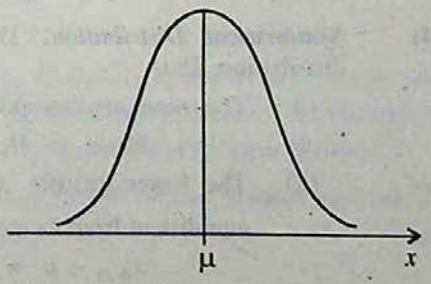


Fig 10.1 Normal distribution

(*i*) Keeping μ constant and decreasing σ causes the density function to become more sharply peaked, thus giving higher probabilities of X being close to μ.

(*ii*) Keeping μ constant and increasing σ causes the density function to flatten, thus giving lower probabilities of X being close to μ.

(*iii*) If σ is held constant and μ is varied, the shape of the density function remains the same with its centre moving to the location of μ.

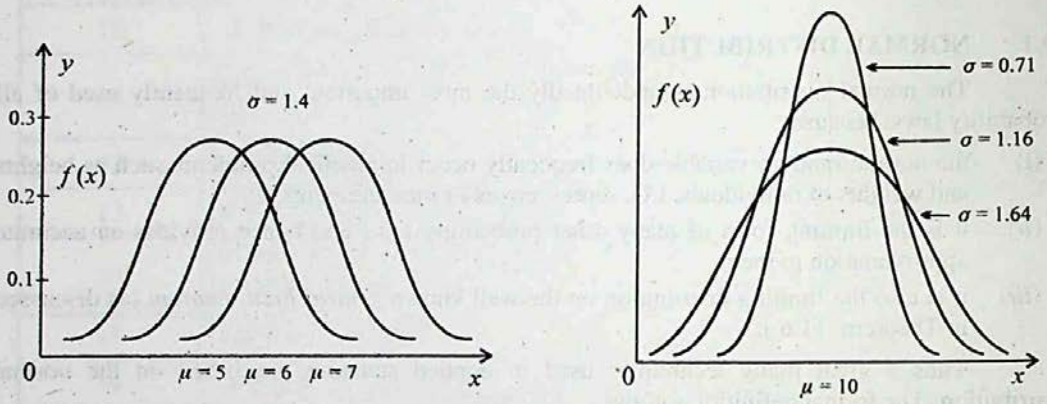


Fig. 10.2 Normal probability density functions

10.1.3 Properties of Normal Distribution. The following are the main properties of normal distribution (or curve).

(1) ***Continuous Distribution.*** The normal probability distribution is a continuous distribution that ranges from $-\infty$ to $+\infty$.

$$R_X = \{ x : -\infty < x < +\infty \}$$

(2) ***Total Probability.*** The total area under the normal curve is unity. That is,

$$P(-\infty < X < +\infty) = 1$$

(3) ***Mode and Maximum Ordinate.*** The normal probability density function is unimodal (single peaked), its mode is μ and its maximum ordinate at $x = \mu$ is

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

(4) ***Symmetrical Distribution.*** The normal probability distribution is a symmetrical distribution. Thus

(*i*) The mean, median and mode coincide at μ.

$$Mean = Median = Mode = \mu$$

(*ii*) The lower quartile $x_{0.25}$ or Q_1 and the upper quartile $x_{0.75}$ or Q_3 are equidistant from its mean μ.

$$x_{0.75} - \mu = \mu - x_{0.25}$$

(*iii*) All odd order moments about mean are zero.

$$\mu_1 = \mu_3 = \mu_5 = \cdots = 0$$

(5) ***Special areas under the curve.*** No matter what values of μ and σ are, the areas under the normal curve remain in fixed proportions within a specified number of σ on either side of μ. For example,

(*i*) $P(\mu - \sigma < X < \mu + \sigma) = 0.6827$

(*ii*) $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$

(*iii*) $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

(6) ***Median, Quartiles and Quartile Deviation.*** In a normal probability distribution, the median, the lower and upper quartiles and the quartile deviation are

$$x_{0.5} = \mu, \quad x_{0.25} = \mu - 0.6745\,\sigma, \quad x_{0.75} = \mu + 0.6745\,\sigma,$$

$$Q.D(X) = \frac{x_{0.75} - x_{0.25}}{2} = 0.6745\,\sigma \approx \frac{2}{3}\sigma$$

$$\mu = \frac{x_{0.25} + x_{0.75}}{2}, \qquad \sigma = \frac{x_{0.75} - x_{0.25}}{1.349}$$

(7) ***Variance, Standard Deviation and Mean Deviation.*** In a normal probability distribution, the variance, the standard deviation and the mean deviation are

$$Var(X) = \sigma^2$$

$$S.D(X) = \sigma$$

$$M.D(X) = \sigma\sqrt{\frac{2}{\pi}} = 0.7979\,\sigma \approx \frac{4}{5}\sigma$$

(8) ***Moments and Moment Ratios.*** In a normal probability distribution, the first four moments about mean and the moment ratios are

$$\mu_1 = 0, \quad \mu_2 = \sigma^2, \quad \mu_3 = 0, \quad \mu_4 = 3\sigma^4$$

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{0}{(\sigma^2)^3} = 0, \qquad \beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3\sigma^4}{(\sigma^2)^2} = 3$

(9) ***Points of Inflexion.*** The points of inflexion of the normal probability density function are equidistant from mean μ, they are at $x = \mu - \sigma$ and $x = \mu + \sigma$. The normal probability distribution is a bell shaped distribution.

(10) ***Asymptotic Curve.*** The normal curve is *asymptotic to the x-axis*, that is, as $|x|$ grows without bound, the curve gets closer and closer to the *x-asis* but always stays above it. The curve has a single peak in the middle and tapers off gradually at both ends and never meets the *x-axis*.

(11) ***Reproductive Property.*** If X_1 and $X2$ are two independent normal random variables having distributinos $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_1^2)$ respectively, then

their sum $X_1 + X_2$ is also a normal random variable having the distribution

$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

10.1.4 Normal Cumulative Distribution Function. The cumulative distribution function for the normal random variable X is

$$F(x) = P(X \le x) = P(-\infty < X \le x)$$

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

Unfortunately, this integration cannot be carried out in the closed form. Numerical techniques could be used to evaluate the integral for specific values of μ and σ. The various possible values of μ and σ result in a family of unlimited number of different normal distributions. It is, thus, necessary to tabulate the standard normal cumulative distribution function, that can be used to evaluate the cumulative distribution function for a normal random variable with any mean μ and any standard deviation σ.

10.2 STANDARD NORMAL RANDOM VARIABLE

A normal random variable X with mean μ and standard deviation σ can easily be transformed into a standard normal random variable Z by the transformation

$$Z = \frac{X - \mu}{\sigma}$$

which has mean 0 and variance 1.

10.2.1 Standard Normal Distribution. If the random variable X has a normal distribution with mean μ and variance σ^2, then the random variable $Z = (X - \mu)/\sigma$ has a standard normal distribution with mean 0 and variance 1.

***Theorem* 10.1** *If* $X \sim N(\mu, \sigma^2)$ *and* $Z = (X - \mu)/\sigma$, *then*

$$Z \sim N(0, 1)$$

Since the standard normal probability density function and cumulative distribution function are of such importance, we shall use ***special symbols*** for them.

10.2.2 Standard Normal Probability Density Function. The probability density function of the standard normal variable Z, denoted by $\varphi(z)$, is given as

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{for } -\infty < z < \infty$$

Thus $\varphi(z)$ is the value of standard normal probability density function at $Z = z$. Therefore, $\varphi(z)$ is called as the ordinate of the standard normal curve at $Z = z$. The ordinates $\varphi(z)$ have been tabulated for various values of z in Table 7.

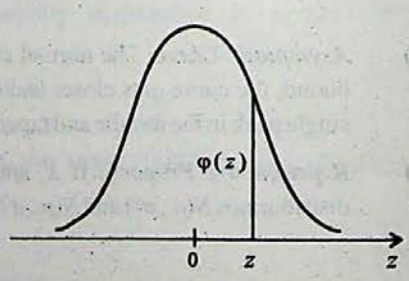


Fig. 10.3 Ordinate of standard normal probability density function at $Z = z$

***Theorem* 10.2** If $Z \sim N(0, 1)$, then

$$\varphi(-z) = \varphi(z)$$

10.2.3 Standard Normal Cumulative Distribution Function. The cumulative distribution function of the standard normal variable Z, denoted by $\Phi(z)$, is given as

$$\Phi(z) = P(Z \le z) = P(-\infty < Z \le z)$$

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

Thus $\Phi(z)$ is the cumulative probability up to $Z = z$ in the standard normal distribution. The cumulative probabilities $\Phi(z)$ have been tabulated for various values of z in Table 9. Note that

$$\Phi(-\infty) = 0$$

$$\Phi(+\infty) = 1$$

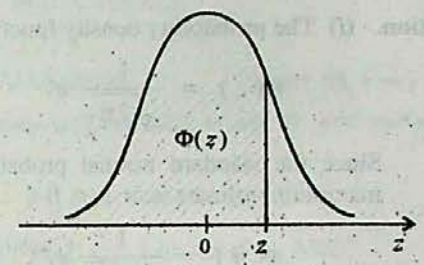


Fig. 10.4 Cumulative probability in standard normal distribution up to $Z = z$

***Theorem* 10.3** *If $Z \sim N(0, 1)$ and a, b are any real numbers, then*

(*i*) $P(Z \le a) = \Phi(a)$

(*ii*) $P(Z \ge a) = 1 - \Phi(a)$

(*iii*) $P(a \le Z \le b) = \Phi(b) - \Phi(a)$

***Theorem* 10.4** If $Z \sim N(0, 1)$, *then for any real value* a

(*i*) $\Phi(-a) = 1 - \Phi(a)$

(*ii*) $P(Z \ge a) = \Phi(-a)$

(*iii*) $P(|Z| \le a) = 2\Phi(a) - 1$

(*iv*) $P(|Z| \ge a) = 2\Phi(-a)$

10.2.4 Inverse Standard Normal Cumulative Distribution Function. The inverse standard normal cumulative distribution function determines a value z corresponding to a given value of the cumulative probability. Suppose that cumulative probability at $Z = z$ is p, then we have

$$\Phi(z) = P(Z \le z) = p$$

$$\Phi^{-1}(p) = z$$

The values of $\Phi^{-1}(p)$ have been tabulated for various values of cumulative probability p in Table 10. For example,

$$\Phi(1.96) = 0.975$$

$$\Phi^{-1}(0.975) = 1.960$$

10.2.5 Use of the Standard Normal Tables. We now show how the tables of the standard normal distribution are used illustrating their direct or inverse use.

***Example* 10.1**

(*i*) *Write down the equation of the standard normal distribution.*

(*ii*) *Find the value of maximum ordinate of the standard normal curve correct to four places of decimal.*

(*iii*) *Verify that the ordinates of the standard normal curve at* $z = 1.27$ *and* $z = -1.27$ *are equal.*

(*iv*) *Find the value* z *when the ordinate at* z *is* 0.12001.

Solution. (*i*) The probability density function of standard normal random variable Z is

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

(*ii*) Since the standard normal probability density function is symmetric about zero, its maximum ordinate is at $z = 0$

$$\varphi(0) = \frac{1}{\sqrt{2\pi}} e^{-(0)^2/2} = \frac{1}{\sqrt{2\pi}} = 0.3989$$

(*iii*) Either calculating directly or using the Table 7, we have

$$\varphi(1.27) = \frac{1}{\sqrt{2\pi}} e^{-(1.27)^2/2} = 0.17810$$

$$\varphi(-1.27) = \frac{1}{\sqrt{2\pi}} e^{-(-1.27)^2/2} = 0.17810$$

Note $\varphi(-1.27) = \varphi(1.27)$

Alternately,

$$\varphi(-1.27) = \frac{1}{\sqrt{2\pi}} e^{-(-1.27)^2/2} = \frac{1}{\sqrt{2\pi}} e^{-(1.27)^2/2} = \varphi(-1.27)$$

(*iv*) By the inverse use of Table 7 and the fact that $\varphi(-z) = \varphi(z)$, we have

$$\varphi(z) = 0.12001$$

$$z = \varphi^{-1}(0.12001) = \pm 1.55$$

***Example* 10.2** *If* Z *is a standard normal random variable with mean* 0 *and variance* 1, *then find*

(*i*) $P(Z < -1.96)$ (*ii*) $P(Z > 1.26)$

(*iii*) $P(-1.96 < Z < 1.96)$ (*iv*) $P(-\infty < Z < 2.12)$

(*v*) $P(-2.72 < Z < \infty)$

Solution. From the definition of standard normal cumulative distribution function, we have

(*i*) $P(Z < -1.96) = \Phi(-1.96)$

$= 0.02500$ (From Table 9)

(*ii*) $P(Z > 1.26) = 1 - P(Z < 1.26)$

$= 1 - \Phi(1.26) = 1 - 0.89617 = 0.10383$

(*iii*) $P(-1.96 < Z < 1.96) = P(Z < 1.96) - P(Z < -1.96)$

$= \Phi(1.96) - \Phi(-1.96)$

$= 0.97500 - 0.02500 = 0.95$

(*vi*) $P(-\infty < Z < 2.12) = P(Z < 2.12) - P(Z < -\infty)$

$= \Phi(2.12) - \Phi(-\infty)$

$= 0.98300 - 0 = 0.98300$ { since $\Phi(-\infty) = 0$ }

(*v*) $P(-2.72 < Z < \infty) = P(Z < \infty) - P(Z < -2.72)$

$= \Phi(\infty) - \Phi(-2.72)$

$= 1 - 0.00326 = 0.99674$ { since $\Phi(+\infty) = 1$ }

***Example* 10.3** *If Z is a standard normal random variable with mean 0 and variance 1, then find*

(*i*) $P(Z < 1.282)$ (*ii*) $P(|Z| < 1.64)$

(*iii*) $P(|Z| > 2.37)$ (*iv*) $P(Z < -1.64 \text{ or } Z > 2.32)$

Solution. From the definition of standard normal cumulative distribution function, we have

(*i*) $P(Z < 1.282) = \Phi(1.282)$ (By interpolating)

$$= \Phi(1.28) + \frac{1.282 - 1.28}{1.29 - 1.28}\{\Phi(1.29) - \Phi(1.28)\}$$

$= 0.89973 + 0.2(0.90147 - 0.89973) = 0.900078$

(*ii*) $P(|Z| < 1.64) = 2\Phi(1.64) - 1$ { since $P(|Z| < a) = 2\Phi(a) - 1$ }

$= 2(0.94950) - 1 = 0.899$

(*iii*) $P(|Z| > 2.37) = 2\Phi(-2.37)$ { since $P(|Z| > a) = 2\Phi(-a)$ }

$= 2(0.00889) = 0.01778$

(*iv*) $P(Z < -1.64 \text{ or } Z > 2.32) = P(Z < -1.64) + P(Z > 2.32)$

$= P(Z < -1.64) + 1 - P(Z < 2.32)$

$= \Phi(-1.64) + 1 - \Phi(2.32)$

$= 0.05050 + 1 - 0.98983 = 0.06067$

***Example* 10.4** *If $Z \sim N(0, 1)$, then find the value of a such that*

(*i*) $P(Z > a) = 0.868$, (*ii*) $P(|Z| < a) = 0.90$

(*iii*) $P(|Z| > a) = 0.238$ (*iv*) $P(Z < a) = 0.6198$

Solution. We have

(*i*) $P(Z > a) = 0.868$

$P(Z < a) = 1 - 0.868 = 0.132$

$\Phi(a) = 0.132$

$a = \Phi^{-1}(0.132) = -1.117$ { From Table 10 (*a*) }

(ii) $P(|Z| < a) = 0.90$

$2\Phi(a) - 1 = 0.90$ { since $P(|Z| < a) = 2\Phi(a) - 1$ }

$\Phi(a) = 0.95$

$a = \Phi^{-1}(0.95) = 1.645$

(iii) $P(|Z| > a) = 0.238$

$2\Phi(-a) = 0.238$ { since $P(|Z| > a) = 2\Phi(-a)$ }

$\Phi(-a) = 0.119$

$-a = \Phi^{-1}(0.119) = -1.18$

$a = 1.18$

(iv) $P(Z < a) = 0.6198$

$\Phi(a) = 0.6198$

$a = \Phi^{-1}(0.6198)$ (By interpolating)

$$= \Phi^{-1}(0.619) + \frac{0.6198 - 0.619}{0.620 - 0.619}\{\Phi^{-1}(0.620) - \Phi^{-1}(0.619)\}$$

$$= 0.3029 + 0.8(0.3055 - 0.3029) = 0.30498$$

10.2.6 Quantiles of Standard Normal Distribution. Let $0 < p < 1$, then the p-th quantile or $(100\,p)$-th percentile of the distribution of standard normal random variable Z is a value z_p such that

$P(Z \le z_p) = p$

$\Phi(z_p) = p$

$z_p = \Phi^{-1}(p)$

Therefore, for the 95-th percentile $z_{0.95}$ of standard normal random variable Z, we have

$P(Z \le z_{0.95}) = 0.95$

$\Phi(z_{0.95}) = 0.95$

$z_{0.95} = \Phi^{-1}(0.95) = 1.645$ { From Table 10 (*a*) }

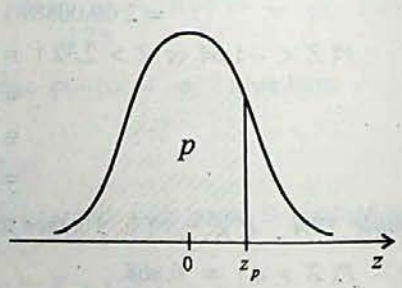


Fig. 10.5 The p-th quantile of standard normal distribution

***Example* 10.5** *If Z is a standard normal random variable, then find the lower and upper quartiles, the inter quartile range, the quartile deviation and the 70-th percentile of the distribution of Z.*

Solution.

For the first quartile $z_{0.25}$, we have

$$P(Z \leq z_{0.25}) = 0.25$$
$$\Phi(z_{0.25}) = 0.25$$
$$z_{0.25} = \Phi^{-1}(0.25)$$
$$z_{0.25} = -0.6745$$

{ From Table 10 (*a*) }

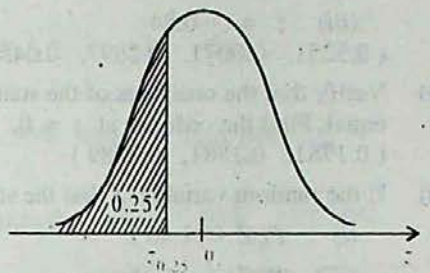


Fig 10.6 First quartile of standard normal distribution

For the third quartile $z_{0.75}$, we have

$$P(Z \leq z_{0.75}) = 0.75$$
$$\Phi(z_{0.75}) = 0.75$$
$$z_{0.75} = \Phi^{-1}(0.75)$$
$$z_{0.75} = 0.6745$$

{ From Table 10 (*a*) }

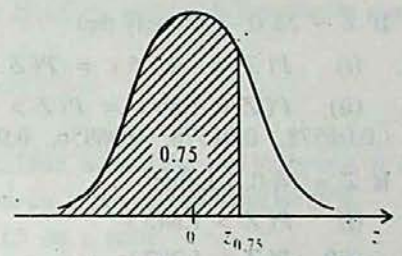


Fig 10.7 Third quartile of standard normal distribution

The inter quartile range (*I. Q. R*) and the quartile deviation (*Q. D*) are

$$I.Q.R = z_{0.75} - z_{0.25} = 0.6745 - (-0.6745) = 1.349$$

$$Q.D = \frac{z_{0.75} - z_{0.25}}{2} = \frac{0.6745 - (-0.6745)}{2} = 0.6745$$

For the 70-th percentile $z_{0.70}$, we have

$$P(Z \leq z_{0.70}) = 0.70$$
$$\Phi(z_{0.70}) = 0.70$$
$$z_{0.70} = \Phi^{-1}(0.70)$$
$$z_{0.70} = 0.5244$$

{ From Table 10 (*a*) }

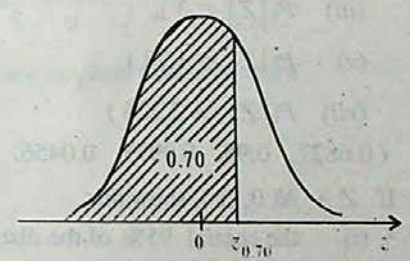


Fig 10.8 Seventieth percentile of standard normal distribution

Exercise 10.1

1. (*a*) Define the normal probability density function and the normal cumulative distribution function. Give the equation of the normal curve with mean μ and standard deviation σ.

(*b*) Define the standard normal probability density function and the standard normal cumulative distribution function. Give the equation of the normal curve with mean 0 and standard deviation 1.

2. (*a*) Find the ordinates of the standard normal curve at

(*i*) $z = 0.64$, (*ii*) $z = 2.84$,

(*iii*) $z = -0.84$ (*iv*) $z = -2.08$

(0.3251, 0.0071, 0.2897, 0.0459)

(*b*) Verify that the ordinates of the standard normal curve at $z = 1.27$ and $z = -1.27$ are equal. Find the ordinate at $z = 0$.

(0.1781, 0.1781, 0.3989)

3. (*a*) If the random variable Z has the standard normal distribution, find

(*i*) $P(Z < 1.46)$ (*ii*) $P(Z > 2.58)$

(*iii*) $P(Z < -1.48)$ (*iv*) $P(Z > -1.96)$

(*v*) $P(0.56 < Z < 1.99)$ (*vi*) $P(-1.32 < Z < 1.65)$

(0.92785, 0.00494, 0.06944, 0.97500, 0.2644, 0.8571)

(*b*) If $Z \sim N(0, 1)$, verify that

(*i*) $P(Z < -2.15) = P(Z > 2.15)$

(*ii*) $P(Z < 1.86) = P(Z > -1.86)$

(0.01578, 0.01578, 0.96856, 0.96856)

4. (*a*) If $Z \sim N(0, 1)$, find

(*i*) $P(Z > 1.645)$ (*ii*) $P(Z < -1.645)$

(*iii*) $P(Z > 1.282)$ (*iv*) $P(Z > 1.96)$

(*v*) $P(Z > 2.576)$ (*vi*) $P(Z > 2.326)$

(*vii*) $P(Z > 2.808)$ (*viii*) $P(Z < -1.96)$

(0.05, 0.05, 0.0999, 0.025, 0.005, 0.01, 0.0025, 0.025)

(*b*) If $Z \sim N(0, 1)$, find

(*i*) $P(|Z| < 1)$ (*ii*) $P(|Z| < 1.96)$

(*iii*) $P(|Z| < 3)$, (*iv*) $P(|Z| > 2)$

(*v*) $P(|Z| < 1.78)$ (*vi*) $P(|Z| < 1.645)$

(*vii*) $P(|Z| > 2.326)$ (*viii*) $P(Z < -1.97 \text{ or } Z > 2.5)$

(0.6827, 0.95, 0.9973, 0.0456, 0.925, 0.9, 0.02, 0.03063)

(*c*) If $Z \sim N(0, 1)$, show that

(*i*) the central 95% of the distribution lies between ± 1.96, *i. e.*,

$$P(-1.96 < Z < 1.96) = 0.95,$$

(*ii*) the central 99% of the distribution lies between ± 2.576, *i. e.*,

$$P(-2.576 < Z < 2.576) = 0.99.$$

5. (*a*) If $Z \sim N(0, 1)$, find a if

(*i*) $P(Z < a) = 0.325$ (*ii*) $P(Z > a) = 0.025$

(*iii*) $P(|Z| < a) = 0.9$ (*iv*) $P(|Z| > a) = 0.097$

(−0.4538, 1.960, 1.645, 1.66)

(*b*) In a standard normal distribution, find

(*i*) a point that has 97.5% area below it, *i. e.*, $z_{0.975}$;

(*ii*) a point that has 97.5% area above it, *i. e.*, $z_{0.025}$;

(*iii*) two such points that contain central 90% area *i. e.*, $z_{0.05}$ and $z_{0.95}$.

(1.96, − 1.96, − 1.645, 1.645)

6. (*a*) If $Z \sim N(0, 1)$, find a if $P(|Z| < a)$ takes the value (*i*) 80% (*ii*) 99%.

(1.282, 2.576)

(*b*) If $Z \sim N(0, 1)$, find a if $P(|Z| > a)$ takes the value (*i*) 5% (*ii*) 2%.

(1.96, 2.326)

7. (*a*) Find the median, the lower and the upper quartiles, and the inter-quartile range for a standard normal random variable Z.

(*b*) In a standard normal distribution,

(*i*) what is the value of mode,

(*ii*) the area to the right of $z = 1$ is 0.1587, what is the area to the left of $z = 1$?

(*iii*) find two points on z scale such that the area between them is 80%,

(*iv*) find the area between − 1.5 and 2.5 on z scale.

(0; 0.8413; − 1.28, 1.28; 0.9270)

10.2.7 Use of the Standard Normal Tables for Any Normal Distribution. We now show how the tables of the standard normal random variable Z can be used for any normal random variable X where $X \sim N(\mu, \sigma^2)$.

***Theorem* 10.5** *If* $X \sim N(\mu, \sigma^2)$, *then*

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

***Theorem* 10.6** *If* $X \sim N(\mu, \sigma^2)$ *and* a, b *are any real numbers, then*

(*i*) $P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$

(*ii*) $P(X \geq a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$

(*iii*) $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

***Theorem* 10.7** *If* $X \sim N(\mu, \sigma^2)$ *and* a *is any real number, then*

$$f(a) = \frac{1}{\sigma}\varphi\left(\frac{a-\mu}{\sigma}\right)$$

***Example* 10.6** *If X is a normal random variable with $\mu = 40$ and $\sigma = 5$, write down its probability density function. Find the ordinate of its normal curve at $x = 42.5$. Also find its maximum ordinate.*

Solution. We have

$$f(a) = \frac{1}{\sigma}\varphi\left(\frac{a-\mu}{\sigma}\right)$$

$$f(42.5) = \frac{1}{5}\varphi\left(\frac{42.5-40}{5}\right)$$

$$= \frac{1}{5}\varphi(0.5) = \frac{1}{5}(0.35207) \qquad \text{(From Table 7)}$$

$$= 0.070414$$

Alternately, The probability density function of the normal random variable X with parameters $\mu = 40$ and $\sigma = 5$ is

$$f(x) = \frac{1}{5\sqrt{2\pi}} e^{-(x-40)^2/2(5)^2} = \frac{1}{5\sqrt{2\pi}} e^{-(x-40)^2/50}$$

$$f(42.5) = \frac{1}{5\sqrt{2\pi}} e^{-(42.5-40)^2/50} = 0.0704$$

The maximum ordinate of this normal curve is at $x = \mu = 40$, which is

$$f(40) = \frac{1}{5}\varphi\left(\frac{40-40}{5}\right)$$

$$= \frac{1}{5}\varphi(0) = \frac{1}{5}(0.39894) \qquad \text{(From Table 7)}$$

$$= 0.079788$$

***Example* 10.7** *The scores made by candidates in a certain test are normally distributed with mean 500 and standard deviation 100. What percent of the candidates received scores*

(*i*) *less than* 400, (*ii*) *more than* 700,

(*iii*) *between* 400 *and* 600, (*iv*) *which differ from mean by more than* 150,

(*v*) *if a candidate gets a score of* 680, *what percent of the candidates have higher scores than he?*

Solution. Let X be the score of a candidate, then $\mu = 500$ and $\sigma = 100$.

(*i*) $$P(X < 400) = P\left(\frac{X-\mu}{\sigma} < \frac{400-500}{100}\right)$$

$$= P(Z < -1) = \Phi(-1) = 0.15866 = 15.87\%$$

(*ii*) $$P(X > 700) = P\left(\frac{X-\mu}{\sigma} > \frac{700-500}{100}\right)$$

$$= P(Z > 2) = 1 - P(Z < 2)$$

$$= 1 - \Phi(2) = 1 - 0.97725 = 0.02275 = 2.28\%$$

(iii) $P(400 < X < 600) = P\left(\frac{400-500}{100} < \frac{X-\mu}{\sigma} < \frac{600-500}{100}\right)$

$= P(-1 < Z < 1) = P(Z < 1) - P(Z < -1)$

$= \Phi(1) - \Phi(-1) = 0.84134 - 0.15866$

$= 0.68266 = 68.27\%$

(iv) $P(|X-\mu| > 150) = P\left(\left|\frac{X-\mu}{\sigma}\right| > \frac{150}{100}\right)$

$= P(|Z| > 1.5) = 2\Phi(-1.5)$

$= 2(0.06681) = 0.13362 = 13.362\%$

(v) $P(X > 680) = P\left(\frac{X-\mu}{\sigma} > \frac{680-500}{100}\right)$

$= P(Z > 1.8) = 1 - P(Z < 1.8)$

$= 1 - \Phi(1.8) = 1 - 0.96407 = 0.03593 = 3.59\%$

***Example* 10.8** *Given that the height of college boys is normally distributed with mean 5′-2″ and standard deviation 4″ and that the minimum height required for joining the N.C.C. is 5′-4″. Find the percentage of boys who would be rejected on account of their height.*

Solution. Let X be the height of a college boy, then $\mu = 5'\text{-}2'' = 62$ inches and $\sigma = 4$ inches. The students with heights less than $5'\text{-}4'' = 64$ inches will be rejected for joining *N.C.C.* Then

$$P(X < 64) = P\left(\frac{X-\mu}{\sigma} < \frac{64-62}{4}\right)$$

$$= P(Z < 0.5) = \Phi(0.5) = 0.69146 = 69.15\%$$

***Example* 10.9** *In a normal distribution with mean μ and standard deviation σ find $P(\mu - \sigma \le X \le \mu + \sigma)$.*

Solution. Let X be a normal random variable with mean μ and standard deviation σ, then

$$P(\mu - \sigma \le X \le \mu + \sigma) = \left(\frac{\mu-\sigma-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{\mu+\sigma-\mu}{\sigma}\right)$$

$$= P(-1 \le Z \le 1) = P(Z \le 1) - P(Z \le -1)$$

$$= \Phi(1) - \Phi(-1) = 0.84134 - 0.15866$$

$$= 0.68268$$

***Example* 10.10** *If the diameters of ball bearings are normally distributed with mean 0.6140 inches and standard deviation 0.0025 inches. Determine the percentage of ball bearings with diameters*

(i) *less than 0.608 inches,* (ii) *greater than 0.617 inches,*

(iii) *between 0.610 and 0.618 inches inclusive,* (iv) *equal to 0.615 inches.*

Solution. Let X be the diameter of a ball bearing, then $\mu = 0.6140$ inches and $\sigma = 0.0025$ inches. Considering the measurement errors, we apply continuity correction to the measurements.

(*i*) The diameter smaller than 0.608 inches is in fact the diameter less than 0.6075 inches. Then

$$P(X < 0.6075) = P\left(\frac{X-\mu}{\sigma} < \frac{0.6075 - 0.6140}{0.0025}\right)$$

$$= P(Z < -2.6) = \Phi(-2.6) = 0.00466 = 0.466\%$$

(*ii*) The diameter greater than 0.617 inches is in fact the diameter more than 0.6175 inches. Then

$$P(X > 0.6175) = P\left(\frac{X-\mu}{\sigma} > \frac{0.6175 - 0.6140}{0.0025}\right)$$

$$= P(Z > 1.4) = 1 - P(Z < 1.4) = 1 - \Phi(1.4)$$

$$= 1 - 0.91924 = 0.08076 = 8.076\%$$

(*iii*) The diameter between 0.610 and 0.618 inches inclusive is in fact the diameter between 0.6095 and 0.6185 inches. Then

$$P(0.6095 < X < 0.6185) = P\left(\frac{0.6095 - 0.6140}{0.0025} < \frac{X-\mu}{\sigma} < \frac{0.6185 - 0.6140}{0.0025}\right)$$

$$= P(-1.8 < Z < 1.8) = P(Z < 1.8) - P(Z < -1.8)$$

$$= \Phi(1.8) - \Phi(-1.8) = 0.96407 - 0.03593$$

$$= 0.92814 = 92.814\%$$

(*iv*) The diameter equal to 0.615 inches is in fact the diameter between 0.6145 and 0.6155 inches. Then

$$P(0.6145 < X < 0.6155) = P\left(\frac{0.6145 - 0.6140}{0.0025} < \frac{X-\mu}{\sigma} < \frac{0.6155 - 0.6140}{0.0025}\right)$$

$$= P(0.2 < Z < 0.6) = P(Z < 0.6) - P(Z < 0.2)$$

$$= \Phi(0.6) - \Phi(0.2) = 0.72575 - 0.57926$$

$$= 0.14649 = 14.649\%$$

10.2.8 De-standardizing. Sometimes it is required to find a value of X that corresponds to the standardized value of Z. We use the relation

$$Z = \frac{X-\mu}{\sigma} \quad \Rightarrow \quad X = \mu + \sigma Z$$

10.2.9 Quantiles of a Normal Distribution. Let $0 < p < 1$, then the p-th quantile or $(100\,p)$-th percentile of the distribution of standard normal random variable Z is a value z_p such that

$$P(Z \le z_p) = p$$

$$\Phi(z_p) = p$$

$$z_p = \Phi^{-1}(p)$$

This value z_p of the p-th quantile or $(100\,p)$-th percentile of the standard normal random variable $Z = (X - \mu)/\sigma$ can be de-standardized for determining the p-th quantile or $(100\,p)$-th percentile x_p of any normal random variable X with parameters μ and σ by the relation

$$X = \mu + \sigma Z$$

Therefore

$$x_p = \mu + \sigma z_p$$

***Example* 10.11** *If* $X \sim N(50, 25)$, *find the value of* X *which corresponds to a standardized value*

(*i*) -1.4, (*ii*) 0, (*iii*) 1.6

Solution. We have $X \sim N(50, 25)$, then $\mu = 50$ and $\sigma^2 = 25 \Rightarrow \sigma = 5$. Then

$$z = \frac{x - \mu}{\sigma} = \frac{x - 50}{5} \Rightarrow x = 50 + 5z$$

Putting the values (*i*) $z = -1.4$, (*ii*) $z = 0$, (*iii*) $z = 1.6$, we get

(*i*) For $z = -1.4$, we get

$$x = 50 + 5z = 50 + 5(-1.4) = 4.3$$

(*ii*) For $z = 0$, we get

$$x = 50 + 5z = 50 + 5(0) = 50$$

(*iii*) For $z = 1.6$, we get

$$x = 50 + 5z = 50 + 5(1.6) = 58$$

***Example* 10.12** *If* $X \sim N(70, 25)$, *find*

(*i*) *a point that has* 87.9% *of the distribution below it,*

(*ii*) *a point that has* 81.7% *of the distribution above it,*

(*iii*) *two such points between which the central* 70% *of the distribution lies.*

Solution. We have $X \sim N(70, 25)$, then $\mu = 70$ and $\sigma^2 = 25 \Rightarrow \sigma = 5$

(*i*) Let a be the point that has 87.9% area below it. Then

$$P(X < a) = 87.9\% = 0.879$$

$$P\left(\frac{X - \mu}{\sigma} < \frac{a - 70}{5}\right) = 0.879$$

$$P\left(Z < \frac{a - 70}{5}\right) = 0.879$$

$$\Phi\left(\frac{a - 70}{5}\right) = 0.879$$

$$\frac{a - 70}{5} = \Phi^{-1}(0.879) = 1.17 \quad \{\text{From Table 10 } (a)\}$$

$$a = 70 + 5(1.17) = 75.85$$

(ii) Let a be the point with that has 81.7% area above it. Then

$$P(X > a) = 81.7\% = 0.817$$

$$P(X < a) = 1 - 0.817 = 0.183$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{a-70}{5}\right) = 0.183$$

$$P\left(Z < \frac{a-70}{5}\right) = 0.183$$

$$\Phi\left(\frac{a-70}{5}\right) = 0.183$$

$$\frac{a-70}{5} = \Phi^{-1}(0.183) = -0.904$$

$$a = 70 + 5(-0.904) = 65.48$$

(iii) Let a, b be the two points between which 70% area lies. Then

$$P(a < X < b) = 70\% = 0.70$$

But $P(X < a) + P(a < X < b) + P(X > b) = 1$ (Total probability)

$$P(X < a) + 0.70 + P(X > b) = 1$$

$$P(X < a) + P(X > b) = 1 - 0.70 = 0.30$$

By symmetry $P(X < a) = P(X > b) = 0.30/2 = 0.15$, therefore

$$P(X < a) = 0.15$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{a-70}{5}\right) = 0.15$$

$$P\left(Z < \frac{a-70}{5}\right) = 0.15$$

$$\Phi\left(\frac{a-70}{5}\right) = 0.15$$

$$\frac{a-70}{5} = \Phi^{-1}(0.15)$$

$$\frac{a-70}{5} = -1.0364$$

$$a = 70 + 5(-1.0364) = 64.818$$

$$P(X > b) = 0.15$$

$$(X < b) = 1 - 0.15 = 0.85$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{b-70}{5}\right) = 0.85$$

$$P\left(Z < \frac{b-70}{5}\right) = 0.85$$

$$\Phi\left(\frac{b-70}{5}\right) = 0.85$$

$$\frac{b-70}{5} = \Phi^{-1}(0.85)$$

$$\frac{b-70}{5} = 1.0364$$

$$b = 70 + 5(1.0364) = 75.182$$

***Example* 10.13** *If* $X \sim N(24, 16)$, *then find the* 33*-rd percentile.*

Solution. We have $X \sim N(24, 16)$, then $\mu = 24$ and $\sigma^2 = 16 \Rightarrow \sigma = 4$

For the 33-rd percentile $x_{0.33}$ or P_{33}, we have

$$P(X < x_{0.33}) = 0.33$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{x_{0.33}-24}{4}\right) = 0.33$$

$$P\left(Z < \frac{x_{0.33}-24}{4}\right) = 0.33$$

$$\Phi\left(\frac{x_{0.33}-24}{4}\right) = 0.33$$

$$\frac{x_{0.33}-24}{4} = \Phi^{-1}(0.33) = -0.4399$$

$$x_{0.33} = 24 + 4(-0.4399) = 22.24$$

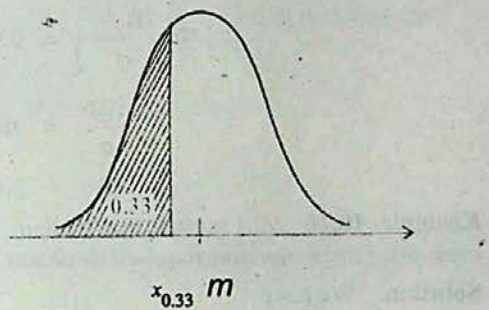


Fig 10.9 Thirty third percentile of the given normal distribution

10.2.10 Finding the values of μ or σ or both.

***Example* 10.14** *If $X \sim N(\mu, 25)$ and $P(X > 69.6) = 0.017$, find the value of the mean, μ.*

Solution. We have $X \sim N(\mu, 25)$, then $\sigma^2 = 25 \Rightarrow \sigma = 5$

$$P(X > 69.6) = 0.017$$

$$P(X < 69.6) = 1 - 0.017 = 0.983$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{69.6-\mu}{5}\right) = 0.983$$

$$P\left(Z < \frac{69.6-\mu}{5}\right) = 0.983$$

$$\Phi\left(\frac{69.6-\mu}{5}\right) = 0.983$$

$$\frac{69.6-\mu}{5} = \Phi^{-1}(0.983) = 2.120$$

$$\mu = 69.6 - 5(2.12) = 59$$

***Example* 10.15** *If $X \sim N(50, \sigma^2)$ and $P(X < 60.6) = 0.983$, find the value of the standard deviation, σ.*

Solution. We have $X \sim N(50, \sigma^2)$, then $\mu = 50$

$$P(X < 60.6) = 0.983$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{60.6-50}{\sigma}\right) = 0.983$$

$$P\left(Z < \frac{10.6}{\sigma}\right) = 0.983$$

$$\Phi\left(\frac{10.6}{\sigma}\right) = 0.983$$

$$\frac{10.6}{\sigma} = \Phi^{-1}(0.983) = 2.120$$

$$10.6 = 2.120\,\sigma \qquad \Rightarrow \qquad \sigma = 5$$

***Example* 10.16** *In a normal distribution* 33% *of the values are under* 48 *and* 12.3% *are over* 60. *Find mean and standard deviation of the distribution.*

Solution. We have

$$P(X < 48) = 33\% = 0.33$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{48-\mu}{\sigma}\right) = 0.33$$

$$P\left(Z < \frac{48-\mu}{\sigma}\right) = 0.33$$

$$\Phi\left(\frac{48-\mu}{\sigma}\right) = 0.33$$

$$\frac{48-\mu}{\sigma} = \Phi^{-1}(0.33) = -0.4399$$

$$48 - \mu = -0.4399\,\sigma \quad \ldots\ldots(i)$$

$$P(X > 60) = 12.3\% = 0.123$$

$$P(X < 60) = 1 - 0.123 = 0.877$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{60-\mu}{\sigma}\right) = 0.877$$

$$P\left(Z < \frac{60-\mu}{\sigma}\right) = 0.877$$

$$\Phi\left(\frac{60-\mu}{\sigma}\right) = 0.877$$

$$\frac{60-\mu}{\sigma} = \Phi^{-1}(0.877) = 1.1601$$

$$60 - \mu = 1.1601\,\sigma \quad \ldots\ldots(ii)$$

Subtracting (i) from (ii), we get

$$\begin{array}{rcl} 60 - \mu &=& 1.1601\,\sigma \\ 48 - \mu &=& -0.4399\,\sigma \\ - \quad + & & + \\ \hline 12 &=& 1.6\,\sigma \end{array} \qquad \Rightarrow \qquad \sigma = 7.5$$

Putting this value of σ in (ii), we have

$$60 - \mu = 1.1601\,(7.5) \qquad \Rightarrow \qquad \mu = 51.3$$

***Example* 10.17** *If X is a normal random variable with parameters $\mu = 50$ and $\sigma = 10$. Find its mean, median, mode, lower and upper quartiles, quartile deviation, mean deviation, variance, standard deviation, first four moments about mean, moment ratios and moment coefficient of skewness.*

Solution. We have $\mu = 50, \quad \sigma = 10$

The mean, median, mode, lower and upper quartiles of the distribution are

$$\text{Mean} = \mu = 50, \quad \text{Median: } x_{0.5} = \mu = 50, \quad \text{Mode} = \mu = 50$$

$$x_{0.25} = \mu - 0.6745\,\sigma = 50 - 0.6745\,(10) = 43.255$$

$$x_{0.75} = \mu + 0.6745\,\sigma = 50 + 0.6745\,(10) = 56.745$$

The quartile deviation, mean deviation, variance and standard deviation of the distribution are

$$Q.\,D(X) = 0.6745\,\sigma = 0.6745\,(10) = 6.745$$

$$M.\,D(X) = 0.7979\,\sigma = 0.7979\,(10) = 7.979$$

$$Var(X) = \sigma^2 = (10)^2 = 100$$

$$S.\,D(X) = \sigma = \sqrt{100} = 10$$

The first four moments about mean, moment ratios and moment coefficient of skewness of the distribution are

$$\mu_1 = 0, \qquad \mu_2 = \sigma^2 = (10)^2 = 100,$$

$$\mu_3 = 0, \qquad \mu_4 = 3\sigma^4 = 3(10)^4 = 30000$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(100)^3} = 0, \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{30000}{(100)^2} = 3$$

$$Sk = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{0}{\sqrt{(100)^3}} = 0$$

***Example* 10.18** *In a normal distribution lower and upper quartiles are* 28 *and* 55 *respectively. Find mean and standard deviation of the normal distribution.*

Solution. We have $x_{0.25} = 28$ and $x_{0.75} = 55$. Then

$$\mu = \frac{x_{0.25} + x_{0.75}}{2} = \frac{28 + 55}{2} = 41.5$$

$$\sigma = \frac{x_{0.75} - x_{0.25}}{1.349} = \frac{55 - 28}{1.349} = 20$$

10.2.11 Normal Distribution as a Limit of a Frequency Distribution of a Continuous Variable. A normal curve serves a good approximation not only for a histogram obtained from the binomial distribution but for many other histograms of observed frequency distributions of continuous random variables as well. Frequently a histogram of an observed frequency distribution with mean $\bar{x}$ and standard deviation s is well approximated for a large number of observations $n = \Sigma f_i$ by the normal curve whose equation is given by

$$f(x) = \frac{nh}{s\sqrt{2\pi}}\, e^{-(x-\bar{x})^2/2s^2}$$

where h is the common class interval in the grouped frequency distribution. The smaller the value of h, the better the approximation will be.

Exercise 10.2

1. (*a*) If X is a normal random variable with $\mu = 24$ and $\sigma = 4$, write down its probability density function. Find the ordinate of its normal curve at $x = 21$. Also find its maximum ordinate.
(0.075285, 0.099735)

(*b*) Suppose that during periods of transcendental meditation, the reduction of a person's oxygen consumption is a random variable have a normal distribution with $\mu = 37.6$ *c. c* per minute and $\sigma = 4.6$ *c. c* per minute. Find the probabilities that during a period of transcendental meditation a person's oxygen will be reduced by

(*i*) at most 35.0 *c. c* per minute,

(*ii*) at least 44.5 *c. c* per minute,

(*iii*) any where from 30.0 to 40.0 *c. c.* per minute.

(0.28604, 0.06681, 0.64900)

(*c*) Let $X \sim N(20, 25)$, find the area under the normal curve

(*i*) below 30,

(*ii*) above 30,

(*iii*) between 30 and 42

(0.97725, 0.02275, 0.02274)

2. (*a*) Suppose that it is know that IQ's for adult Pakistanis are normally distributed with $\mu = 100$ and $\sigma = 10$. If an individual with an IQ of 130 or above is classified genius, what is the probability that a random selection yields a genius?

(0.00135)

(*b*) The mean inside diameter of a sample of 250 washers produced by a machine is 5.05 *mm* and the standard deviation is 0.05 *mm*. The purpose for which these washers are intended allows a maximum tolerance in the diameter of 4.95 *mm* to 5.10 *mm*, otherwise the washers are considered defective. Determine the percentage of defective washers produced by the machine assuming the diameters are normally distributed

(18.14%)

3. (*a*) The length of life for a washing machine is approximately normally distributed, with a mean of 3.5 years and a standard deviation of 1.0 years. If this type of washing machine is guaranteed for 12 months, what percentage of the sales will require replacement?

(0.62%)

(*b*) Assume that the time X required for a runner to run a mile is a normal random variable with parameters $\mu = 4$ minutes 1 second and $\sigma = 2$ seconds. What is the probability that this athlete will run the mile

(*i*) in less than 4 minutes,

(*ii*) in more than 3 minutes 55 seconds.

(0.3085, 0.9987)

(*c*) Assume that the distance X that a particular athlete will be able to put a shot (on his first try) is normally distributed with parameters $\mu = 50$ feet and $\sigma = 5$ feet. Compute the probability that he tosses it not less than 55 feet and the probability that his toss travels between 50 feet and 60 feet.

(0.1587, 0.4773)

(*d*) If X is normally distributed with parameters μ and σ, find the area under the curve between

(*i*) $(\mu - \sigma)$ and $(\mu + \sigma)$,

(*ii*) μ and $(\mu + 2\sigma)$.

(0.68268, 0.47725)

4. **(*a*)** The heights of boys at a particular age follow a normal distribution with mean 150.3 *cm* and standard deviation 5.0 *cm*. Find the probability that a boy picked at random from this age group has height

(*i*) less than 153 *cm*,

(*ii*) more than 145 *cm*,

(*iii*) between 146 *cm* and 152 *cm*.

(0.67003, 0.83147, 0.5015)

(*b*) Suppose the weekly incomes X are normally distributed with mean 10.06 Rs. and variance 2.64 Rs^2., find the probability $P(8 \leq X \leq 12)$. Assume that the incomes are recorded to the nearest rupee.

(0.87569)

5. **(*a*)** Find the value of X which corresponds to a standardized value of -2.05 and 0.86 for each of the following distributions

(*i*) $X \sim N(62.3, 38)$,

(*ii*) $X \sim N(\mu, \sigma^2)$,

(*iii*) $X \sim N(a, b)$.

(49.66, 67.60; $\mu - 2.05\sigma$, $\mu + 0.86\sigma$; $a - 2.05\sqrt{b}$, $a + 0.86\sqrt{b}$)

(*b*) If $X \sim N(100, 64)$, find the value of a such that $P(X < a) = 0.95$.

(113.16)

(*c*) If $X \sim N(60, 25)$, find the value of a such that $P(X > a) = 0.837$.

(55.09)

6. **(*a*)** In a normal distribution $\mu = 30$ and $\sigma = 5$. Find

(*i*) a point that has 15% area below it,

(*ii*) a point that has 28% area above it,

(*iii*) two points containing middle 95% area,

(24.8; 32.9; 20.2 and 39.8)

(*b*) The time required by a nurse to inject a shot of penicillin has been observed to be normally distributed, with a mean of $\mu = 30$ seconds and a standard deviation of $\sigma = 10$ seconds. Find the

(*i*) 10-th percentile, *i. e.*, $x_{0.10}$,

(*ii*) 90-th percentile, *i. e.*, $x_{0.90}$.

(17.2 *sec*, 42.8 *sec*)

(*c*) Scores on a national education achievement test are normally distributed with $\mu = 500$ and $\sigma = 100$.

(*i*) What is the 95-th percentile of this distribution,

(*ii*) What are the lower and upper quartiles of this distribution,

(*iii*) If the university decides to accommodate the 40 percent of the students with the highest scores, what is the score that separates the successful applicants with unsuccessful?

(664.5; 432.55, 567.45; 525.33)

7. **(*a*)** The height a high jumper will clear, each time he jumps, is a normal random variable with mean 6 feet and standard deviation 2.4 inches.

(*i*) What is the greatest height he will jump with the probability 0.95?

(*ii*) What is the height he will clear only 10% of the time?

(68.06 inches, 75 inches)

(*b*) Suppose that the amount of vaccine required to immunize human beings against smallpox is normally distributed with $\mu = 0.250$ ounce and $\sigma = 0.040$ ounce. Increasing the dosage increase the chances of successful vaccination. What is the minimum dosage required to produce success in 99 percent of the cases.

(0.34304 *ounces*)

8. **(*a*)** If $X \sim N(70, 25)$, find the value of a such that $P(|X - 70| < a) = 0.8$. Hence find the limits within which the central 80% of the distribution lies.

(6.41, 63.59, 76.41)

(*b*) Bags of flour by a particular machine have masses which are normally distributed with mean 500 g and standard deviation 20 *g*. 2% of the bags are rejected for being underweight and 1% of the bags are rejected for being over weight. Between what range of values should the mass of a bag of flour lie is to accepted?

(458.92, 546.52)

9. **(*a*)** The lengths of items follow a normal distribution with mean μ *cm* and standard deviation 12 *cm*. It is known that 4.78% of the items have a length greater than 92 *cm*. Find the value of mean μ.

(72)

(*b*) The lengths of rods produce in a workshop follow a normal distribution with mean μ and variance 4. If 10% of the rods are less than 17.4 *cm* long. Find the probability that a rod chosen at random will be between 18 and 23 *cm* long.

(0.7725)

10. **(*a*)** Tea is soled in packages marked 750 *g*. The masses of the package are normally distributed with mean 760 *g* standard deviation σ. What is the maximum value of the σ if less than 1% of the packages are under weight?

(4.299)

(*b*) Suppose that the life in hours of an electric tube manufactured by a certain process is normally distributed with parameters $\mu = 160$ hours and σ hours. What is the maximum allowable value for σ, if the life X of a tube is to have probability 0.80 of being between 120 and 200 hours?

(31.21)

11. **(*a*)** Assume that we have a large number of students whose average weight is 150 *lb* and that the weights are normally distributed. If we know that 36.4% of the students have weights between 137 and 163 *lb*. What is the standard deviation of the weights?

(27.47)

(*b*) In a normal distribution $\mu = 40$ and $P(25 < X < 55) = 0.8662$. Find $P(20 < X < 60)$.

(0.9545)

12. **(*a*)** In a normal distribution 31% of item are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

(50, 10)

(*b*) If $X \sim N(\mu, \sigma^2)$ and $P(X < 35) = 0.20$ and $P(35 < X < 45) = 0.65$. Find μ and σ.
(39.5, 5.32)

13. (*a*) Assuming that the number of marks scored by a candidates is normally distributed, find the mean and the standard deviation, if the number of first class students (60% or more marks) is 25, the number of failed students (less than 30% marks) is 90 and the total number of candidates appearing for the examination is 450.
(40.37, 12.32)

(*b*) A marketing organization grades apples into three sizes, small (diameter less than 60 mm), medium (diameter between 60 and 80 mm), and large (diameter more than 80 mm). A certain grower finds that 61% of his crop falls into the small category, and 14% into the large category. Assuming that the distribution of the diameter X of the apples is described by a normal probability density function, calculate the standard deviation and mean of his crop.
(24.97, 53.03)

(*c*) The maximum temperature on June, 1 in a certain locality has been recorded and observed as normally distributed over year. About 15% of the time, it has exceeded $30^\circ C$, and about 5% of the time, it has been less than $20^\circ C$. What is the mean and variance of the data?
($26.13^\circ C$, $13.91^\circ C$)

14. (*a*) A man cuts hazel twigs to make bean poles. He says that a stick is 240 *cm* long. In fact, the length of the stick follows a normal distribution and 10% are of length 250 *cm* or more while 55% have a length over 240 *cm*. Find the probability that a stick picked at random is less than 235 *cm* long.
(0.203)

(*b*) The 90-th percentile of a normal distribution is 50 while the 15-th percentile is 25.
(*i*) Find μ and σ.
(*ii*) What is the value of 40-th percentile.
(36.17, 10.79, 33.44)

15. (*a*) The masses of articles produced in a particular workshop are normally distributed with mean μ and standard deviation σ. The 5% of the articles have a mass greater than 85 *g* and 10% have a mass less than 25 *g*. Find the value of μ and σ, and find the range symmetrical about the mean, within which 75% of the masses lie.
(51.3, 20.5, 26.72, 73.88)

(*b*) In a certain examination, the percentage of passes and distinctions were 80 and 10 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively, assume the distribution of marks to be normal.
(53.87, 16.48)

16. (*a*) What is the importance of normal distribution in statistical theory? Describe its properties.

(*b*) Suppose that X is normally distributed with $\mu = 25$ and $\sigma = 5$. Find
(*i*) the lower and upper quartiles,

(*ii*) the median,

(*iii*) the mean deviation.

(21.6, 28.4, 25, 4)

(*c*) In a normal distribution, the lower and upper quartiles are respectively 8 and 17. Find mean and standard deviation of the normal distribution.
(12.5, 6.67)

(*d*) The continuous random variable X is normally distributed with mean μ and standard deviation σ. Given that $P(X < 53) = 0.04$ and $P(X < 65) = 0.97$. Find the inter-quartile range of the distribution.
(4.46)

17. (*a*) The value of second moment about the mean in a normal distribution is 4. Find the third and the fourth moments about the mean in the distribution.
(0, 48)

(*b*) Find the proportion of the area under the normal curve included between the limits $\mu \pm \sigma$, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$ where μ and σ denote the mean and the standard deviation.
(0.6827, 0.9545, 0.9973)

(*c*) If X is a random variable with distribution $N(12.5, 9)$ and the random variable $Y = g(X) = 2X + 5$. Find $P(-\infty \leq Y \leq 21)$ and $P(45 \leq Y \leq \infty)$.
(0.06681, 0.00621)

◆

Exercise 10.3
Objective Questions

1. Fill in the blanks.

(*i*) The normal distribution is a ———— distribution that ranges from $-\infty$ to ∞. (*continuous*)

(*ii*) The value of the parameter σ of a normal distribution is always ————. (*positive*)

(*iii*) The normal distribution is a bell shaped ———— distribution. (*symmetrical*)

(*iv*) If $X \sim N(50, 25)$, then σ = ————. (5)

(*v*) The maximum ordinate of the standard normal curve is at Z = ————. (0)

(*vi*) In a standard normal distribution, if $P(Z < z_{0.975}) = 0.975$, then $z_{0.975}$ = ————. (1.96)

(*vii*) The maximum ordinate of a normal curve is at X = ————. (μ)

(*viii*) The total area under a normal curve is ————. (*unity*)

(*ix*) The ———— of a normal distribution corresponds to $z = 0$ in the standard normal distribution. (*mean*)

(*x*) In a normal distribution, the mean, median and mode are ———— . *(equal)*

2. Fill in the blanks.

(*i*) In a normal distribution, ———— $= \mu - 0.6745\,\sigma$. (Q_1)

(*ii*) In a normal distribution, ———— $= \mu + 0.6745\,\sigma$. (Q_3)

(*iii*) In a normal distribution, $QD \equiv$ ———— σ. $(2/3)$

(*iv*) In a normal distribution, $MD \equiv$ ———— σ. $(4/5)$

(*v*) The limits $\mu \pm 0.6745\,\sigma$ include ———— percent area under the normal curve. (50)

(*vi*) The limits $\mu \pm \sigma$ include ———— percent area under the normal curve. (68.27)

(*vii*) The limits $\mu \pm 2\,\sigma$ include ———— percent area under the normal curve. (95.45)

(*viii*) The limits $\mu \pm 3\,\sigma$ include ———— percent area under the normal curve. (99.73)

(*ix*) In a normal distribution, all odd ordered moments about mean are ————. *(zero)*

(*x*) In a normal distribution, $\beta_1 = 0$ and β_2 = ————. (3)

(*xi*) The normal distribution is neither platykurtic nor leptokurtic but ————. *(mesokurtic)*

(*xii*) The points of inflexion of a normal curve are ———— from mean. (equidistant)

3. Mark off the statements as true or false.

(*i*) Normal distribution has two parameters namely μ and σ^2. *(true)*

(*ii*) If X is normally distributed with mean μ and variance σ^2 then it is denoted by $X \sim N(\mu, \sigma^2)$. *(true)*

(*iii*) The standard normal distribution has mean 0 and variance 1. *(true)*

(*iv*) The maximum ordinate of a standard normal curve is at $Z = 1$. *(false)*

(*v*) The standard normal distribution is symmetrical about $Z = 0$. *(true)*

(*vi*) In a standard normal distribution, if $P(Z < z_{0.025}) = 0.025$, then $z_{0.025} = -1.96$. *(true)*

(*vii*) In a standard normal distribution, if $P(Z < z_{0.975}) = 0.975$, then $z_{0.975} = 1.96$. *(true)*

(*viii*) In a standard normal distribution, if $P(|Z| < a) = 0.95$, then $a = 1.96$. *(true)*

(*ix*) The normal curve has maximum ordinate at $X = 0$. *(false)*

4. Mark off the following statements as false or true.

(*i*) The shape of a normal distribution depends upon its parameters namely μ and σ. (*true*)

(*ii*) The parameter σ controls the relative flatness of the normal curve. (*true*)

(*iii*) The normal distribution is a bell-shaped symmetrical distribution. (*true*)

(*iv*) In a normal distribution, the mean, median and mode are equal. (*true*)

(*v*) In a normal distribution,

$$Q_1 = \mu - 0.6745\,\sigma \text{ and } Q_3 = \mu + 0.6745\,\sigma.$$ (*true*)

(*vi*) In a normal distribution, mean and variance are always equal. (*false*)

(*vii*) The expected value of a normal distribution is μ. (*true*)

(*viii*) The standard deviation of a normal distribution is σ. (*true*)

(*ix*) The quartile deviation of a normal distribution is $0.6745\,\sigma$. (*true*)

(*x*) The mean deviation of a normal distribution is $0.7979\,\sigma$. (*true*)

(*xi*) The two points containing the middle 95.45% area under a normal curve are $\mu \pm \sigma$. (*false*)

(*xii*) In a normal distribution, all even ordered moments about mean are zero. (*false*)

(*xiii*) In a normal distribution, all odd ordered moments about mean are zero. (*true*)

(*xiv*) In a normal distribution, $\beta_1 = 0$ and $\beta_2 = 3$. (*true*)

(*xv*) The points of inflexion of the normal curve lie at $\mu \pm 2\,\sigma$. (*false*)

(*xvi*) The normal curve gets closer and closer to the *x-axis* but never touches it. (*true*)

11 SAMPLING TECHNIQUES AND SAMPLING DISTRIBUTIONS

11.1 POPULATION (OR UNIVERSE)

A *population* is the totality of the observations made on all the objects (under investigation) possessing some common specific characteristics, which are of particular interest to researchers.

The population is the aggregate of the elements and these elements are the *basic units* that comprise and define a population. The population must be defined in terms of

(*i*) content, (*ii*) unit, (*iii*) extent, (*iv*) time

For instance, the students of first year class at a given college, the characteristic to be investigated may be the score received by each student in a college entrance examination, in a given year. Populations may be finite or infinite.

11.1.1 Finite Population. A population is said to be *finite* if it includes a limited number of elementary units (objects or observations).

Examples of a finite population are: the heights of all the students enrolled at a college in a given year, the wages of all employees of a steel mill in a given year, the amount of money spent by each student in an engineering university in a given academic year, or the grading of items as defective and non-defective that are produced by an industry on a given day.

11.1.2 Infinite Population. A population is said to be *infinite* if it consists of unlimited number of elementary units. At least hypothetically, there is no limit to the number of units it can include.

Examples of an infinite population are: the weights at birth of all human beings, the results obtained by rolling of a die, the lifetimes of all the bulbs produced in a production process that operates indefinitely under given manufacturing conditions.

11.2 SAMPLE

A *sample* is a part of the population which is selected with the expectation that it will represent the characteristics of the population.

11.2.1 Sampling. *Sampling* is a procedure of selecting a representative sample from a given population.

11.2.2 Sample Survey versus Complete Enumeration. The collection of information from a part of the population is called making a *sample survey.* The collection of information from all elements in a population is called taking a *census* or making a *complete enumeration.*

11.2.3 Purposes of Sampling. The two basic purposes of sampling are:

(*i*) To obtain maximum information about the characteristics of the population with minimum cost, time and effort.

(*ii*) To find the reliability of the estimates derived from the sample.

11.2.4 Advantages of Sampling. Following are the main advantages of sampling over a complete census.

(*i*) ***Time Saving***: A sample survey involves lesser amount of time and energy than a complete enumeration both in the execution and the analysis of data. This is a vital consideration when the information is urgently needed as the results from a sample survey are more readily available.

(*ii*) ***Economic***: A sample survey requires less expenses and labour as compared to a complete census because the cost of covering only a fraction will be lower than that of covering the whole population.

(*iii*) ***Accuracy***: A sample survey provides the results which are almost as accurate as those obtained by complete census. A properly designed and carefully executed sample survey will provide even better results.

(*iv*) ***Feasibility***: Sometimes the data are obtained by tests that are destructive. For example, to know the average life of certain type of electric bulb, we shall take a sample of these bulbs and keep them on until they burn out. We cannot think of testing the whole lot. In testing blood of a patient we do not drain the entire blood out of him but examine just a few drops. Sampling may be the only means available for obtaining the desired information when the population is infinite or inaccessible. In such cases complete enumeration would neither be physically possible nor practically feasible.

Whatever be the merits of sampling, it cannot totally replace a complete census. A census is a record of a nation's history and its importance has to be given due acknowledgement.

11.2.5 Limitations of Sampling. If the basic facts of each and every unit in the population are needed, census become indispensable. The sample will not meet such a requirement.

For example, the list of income tax payers is prepared very carefully, the list of voters is prepared to include the name of each and every voter, or an inventory of all goods and stocks is necessary to know the total amount of stocks of a firm.

11.3 SAMPLING DESIGN

A *sampling design* is a procedure or plan for obtaining a sample from a given population prior to collecting any data.

The collection of detailed information is known as *survey*. When a survey is carried out by a sampling design, it is called a *sample survey*. A sample survey should be properly planned and carefully executed in order to avoid inaccuracies.

11.3.1 Sampling Units. *Sampling units* are those basic units of the population in terms of which the sample design is planned.

The sampling units must be distinct and exhaustive, *i. e.*, they must make up the whole population and they must not be overlapping. Sometimes the sampling unit is obvious, as in a population of students or in a population of light bulbs. Sometimes there is a choice of sampling unit. In sampling an agricultural crop, the sampling unit might be a field, a farm or an area of land

whose shape and area is at our disposal. In sampling a human population, the unit might be an individual person, the household or all the persons living in a block.

11.3.2 Sampling Frame. A *sampling frame* is a complete list of the sampling units.

For example, a complete list of all the students in a college on May 10, 1995, is the frame. A complete list of all households in a city is an other example of the frame.

11.3.3 Types of Sampling Designs. Meaningfulness of estimates, obtained from a sample, depends upon the methods of selecting a sample. Broadly speaking there are two different sampling schemes.

(*i*) Non-probability sampling

(*ii*) Probability sampling

11.4 NON-PROBABILITY (NON-RANDOM) SAMPLING

A *non-probability sampling* is a procedure in which we cannot assign to an element of population the probability of its being included in the sample.

We often make inferences about the population from arbitrary and informal samples. A wheat dealer forms his opinion about a sackful of wheat by examining just a few grains. To say about the quality of rice cooked in a big pot, the cook takes only a spoonful of rice to taste and decide on its quality of cooking. Such arbitrary selections are frequently made in research, in biological and physical sciences.

11.5 PROBABILITY (RANDOM) SAMPLING

A *probability sampling* is a process in which the sample is selected in such a way that every element of a population has a known non-zero (not necessarily equal) probability of being included in the sample.

The advantage of probability sampling is that it provides a measure of precision of the estimates. The underlying principle of a random sample is that personal factor is eliminated in the selection as the investigator does not exercise his discretion in the choice of items. No factor other than chance affects the likelihood of an item being included in or excluded from the sample. A random sample may be taken with or without replacement.

11.5.1 Random Sampling With Replacement. Sampling is said to be *with replacement* from a population (finite or infinite) when the unit selected at random is returned to the population before the next unit is selected. The formal description of the sampling method is as follows.

1. An object is selected from the population in a way that gives all objects in the population an equal chance of being selected.
2. The characteristics level of the object selected is observed, and the object is returned to the population prior to any subsequent selection.
3. For a sample of size n, steps (1) and (2) are performed n times

Thus the number of units available for future drawing is not affected. The population remains the same and a sampling unit might be selected more than once.

11.5.2 Random Sampling Without Replacement. Sampling is said to be *without replacement*, when the sampling unit selected at random is not returned to the population, before the next unit is selected. The formal description of the sampling method is as follows.

1. The first object is selected from the population in a way that gives all objects in the population an equal chance of being selected.

2. The characteristics level of the object selected is observed, but the object is not returned to the population.
3. An object is selected from the remaining objects in the population in a way that gives all the remaining objects an equal chance of being selected, and step (2) is repeated. For a sample of size n, step (3) is performed $(n - 1)$ times.

Thus the number of units remaining after each drawing will be reduced by one. In this case, a sampling unit selected once cannot be selected again for the sample because the selected unit is not replaced.

11.6 SIMPLE RANDOM SAMPLING

Simple random sampling is a procedure of selecting a sample of n units $(n = 1, 2, \cdots, N)$ from the population of N units in such a way that:

(*i*) Every unit available for sampling has an equal probability of being drawn.

(*ii*) Every sample of size n has the same probability of being selected.

A sample drawn by this procedure is called a *simple* or *unrestricted random sample.* A sample containing n elements selected from a population consisting of N elements is called a sample of size n. The simple random sampling is used in the population which is essentially homogeneous in terms of some characteristics relevant to the enquiry. For small populations where the elements are easily identifiable and accessible, simple random sampling may be easy to apply.

***Theorem* 11.1** *If a simple random sample of size n is selected from a finite population of size N, then the number of all possible samples is given as*

$$\textit{No. of possible samples} = N^n, \qquad \textit{if sampling is done with replacement}$$

$$\textit{No. of possible samples} = {}^{N}P_{n}, \qquad \textit{if sampling is done without replacement}$$

Proof. A sample of size n under simple random sampling (with or without replacement) consists of an ordered specification of n elements, namely

(the first chosen, the second chosen, $\cdots$, the n-th chosen.)

Sampling With Replacement. If we use sampling with replacement, the number of units available for each drawing are N.

The first unit of the sample can be selected in N different ways, the second unit of the sample can also be selected in N ways, the third unit of the sample can also be selected in N ways, and so on, the n-th unit of the sample can also be selected in N ways.

Using the multiplicative principle, the number of all possible samples of size n, that could be selected from a finite population of size N, is

$$\begin{aligned}\text{No. of possible samples} &= N \times N \times N \times \cdots \; n \text{ times} \\ &= N^n\end{aligned}$$

A sample of n units constitutes only one arrangement, and there are N^n possible arrangements of n units from a finite population of N units. Each of the N^n possible samples is selected with the same probability

$$\frac{1}{N} \times \frac{1}{N} \times \frac{1}{N} \times \cdots n \text{ times} = \frac{1}{N \times N \times N \times \cdots n \text{ times}} = \frac{1}{N^n}$$

Sampling Without Replacement. If we use sampling without replacement, the number of units remaining after each drawing will be reduced by one

The first unit of the sample can be selected in N different ways, the second unit of the sample can be selected in $(N-1)$ ways, the third unit of the sample can be selected in $(N-2)$ ways, and so on, the n-th unit of the sample can be selected in $(N-n+1)$ ways.

Using the multiplicative principle, the number of all possible samples of size n, that could be selected from a finite population of size N, is

$$\begin{aligned}\text{No. of possible samples} &= N(N-1)(N-2)\cdots\{N-(n-1)\}\\ &= N(N-1)(N-2)\cdots(N-n+1)\\ &= \frac{N(N-1)\cdots(N-n+1)(N-n)\cdots(3)(2)(1)}{(N-n)\cdots(3)(2)(1)}\\ &= \frac{N!}{(N-n)!} = {}^{N}P_{n}\end{aligned}$$

A sample of n units constitutes only one arrangement, and there are ${}^{N}P_{n}$ possible arrangements of n units from a finite population of N units. Each of the ${}^{N}P_{n}$ possible samples is selected with the same probability

$$\frac{1}{N}\times\frac{1}{N-1}\times\cdots\times\frac{1}{N-n+1} = \frac{1}{N(N-1)\cdots(N-n+1)} = \frac{1}{{}^{N}P_{n}}$$

11.6.1 Random Digits. A table of *random digits* consists of a sequence of digits designed to represent the result of a simple random sampling with replacement from a population of digits 0, 1, 2, $\cdots$, 9. In a table of random digits each digit from 0 to 9 is called a random digit, each having the probability of occurrence of $1/10$. *Here random implies that all of these digits have the same probability of occurrence and the occurrence and non-occurrence of any digit is independent of the occurrence and non-occurrence of all other digits.* Table 15 is such a table.

In a random digits table, random digits are normally combined to form numbers of more than one digit. For example, random digits taken in pairs will result in a set of 100 different numbers from 00 to 99, each having a probability of occurrence of $1/100$ and each being independent of other numbers similarly formed. Likewise random digits taken in triples will result in a set of 1000 different numbers from 000 to 999, each having a probability of occurrence of $1/1000$ and each being independent of other numbers similarly formed. Similarly, random digits taken in quadruples will result in 10000 different numbers from 0000 to 9999, each having a probability of occurrence of $1/10000$, and each being independent of other numbers similarly formed.

11.6.2 Selection of Simple Random Sample. A simple random sample can be selected by the following methods.

(i) ***Lottery Method.*** In this method, a distinct and different serial number from 1 to N is assigned to every unit of the population of N units and the number is recorded on a card or a slip of paper. All the numbered slips are then placed in a container, and they are thoroughly mixed. A blind selection is made of the number of slips required to constitute the desired size of the sample. The items corresponding to the slips drawn will constitute the random sample. The selection of items depends entirely on chance. Some lotteries use a rotating wheel in selecting tickets. The wheel has equal segments on its rim, one

for each of the digits 0 through 9. N lottery tickets are numbered from 1 to N. Suppose the tickets have three-digit numbers. A ticket number then would be selected by spinning the wheel thrice and recording the digit which appears at the pointer each time the wheel stops. If the digit sequence is 534, then the ticket number 534 is selected. The lottery method becomes quite cumbersome to use as the size of population becomes large, then an alternative method of selection of a random sample is employed.

(*ii*) ***Using Random Digits.*** In this method a distinct sampling number from 0 to $(N-1)$ is assigned to every unit of the population of N units. A table of random digits is consulted with a randomly selected starting point in the table. The table is read in single digits, in groups of two, three or more according to the number of digits in the sampling number $(N-1)$ assigned to the last unit in the population. Any number greater than $(N-1)$ is discarded. A number appearing second time is also discarded if the sampling is without replacement. Continue the process of selecting the random digits or numbers until the desired sample size is reached.

11.7 STRATIFIED SAMPLING

If the elements in the population are not homogenous, then the population is divided into non-overlapping homogeneous subgroups, called *strata*, and sample is drawn separately from each stratum by simple random sampling. This sample is called *stratified random sample*. The process of dividing a heterogeneous population into homogeneous subgroups is called *stratification*.

The benefit of this method is that if non-overlapping homogeneous subgroups of the population can be identified, then only a relatively small number of observations are needed to ascertain the characteristics of each subgroup. Stratification is used also to improve sample estimates of population characteristics. Stratification is used:

(*i*) to provide an adequate sample for each stratum,

(*ii*) because it can give more precise estimates of population characteristics than other types of samples.

11.8 ERRORS

11.8.1 True Value. By *true value* we mean the value that would be obtained if no errors were made in any way in obtaining the information or computing the characteristic of the population.

True value of the population is possibly obtained only if the exact procedures are used for collecting the correct data, each and every element of the population has been covered and no mistake or even the slightest negligence has happened during the process of data collection and its analysis. It is usually regarded as an unknown constant.

11.8.2 Accuracy. By *accuracy* we refer to the difference between the sample result and the true value. The smaller the difference, the greater will be the accuracy. Accuracy can be increased:

(*i*) By elimination of technical errors.

(*ii*) By increasing the sample size.

11.8.3 Precision. By *precision* we refer to how closely we can reproduce, from a sample, the results which would be obtained if a complete count (census) was taken using the same method of measurement.

11.8.4 Error. The difference between an estimated value and the population true value is called an *error*. Since a sample estimate is used to describe a characteristic of a population. A sample being only a part of a population cannot provide a perfect representation of the

population, no matter how carefully the sample is selected. We may think as to how close will the sample estimate be to the population true value. Generally it is seen that an estimate is rarely equal to the true value. There are two kinds of errors:

(*i*) Sampling (random) errors

(*ii*) Non-sampling (non-random) errors

11.8.5 Sampling Error. A *sampling error* is the difference between the value of a statistic obtained from an observed random sample and the value of corresponding population parameter being estimated.

A sample may not provide a true representation of the population under study, simply because samples represent only a part of a population and thus depend on " the luck of the draw ", even if the sample survey is properly designed and well-implemented. Generally, let T be sample statistic used to estimate the population parameter θ, then the sampling error, denoted by E, is defined as

$$E = T - \theta$$

The value of sampling error reveals the *precision* of the estimate. Smaller the sampling error, the greater will be the precision of the estimate. The sampling errors can be reduced:

(*i*) By increasing the sample size.

(*ii*) By improving the sampling design.

(*iii*) By using the supplementary information.

11.8.6 Non-sampling Errors. The errors that are caused by sampling the wrong population of interest and by response bias, as well as those made by an investigator in collecting analysing and reporting the data, are all classified as *non-sampling* or *non-random errors*. These errors are present in a complete census as well as in a sample survey.

11.8.7 Bias. *Bias* is the difference between the expected value of a statistic and the true value of the parameter being estimated. Let T be the sample statistic used to estimate the parameter θ, then the amount of bias is

$$Bias = E(T) - \theta$$

The bias is positive if $E(T) > \theta$, it is negative if $E(T) < \theta$ and it is zero if $E(T) = \theta$. Bias is a systematic component of error which refers to the long-run tendency of the sample statistic to differ from the parameter in a particular direction. Bias is cumulative and increases with the increase in size of the sample. If proper methods of selection of units in a sample are not followed, the sample results will not be free from bias.

Exercise 11.1

1. (*a*) Explain the terms: Population; Sample; Sampling frame; Sampling unit.

(*b*) Define suitable populations from which the following samples are selected:

(*i*) One thousand homes are called by telephone in the city of Karachi and asked to name the T.V programme that they are now watching

(*ii*) A coin is flipped 53 times and 32 heads are recorded.

(*iii*) Two hundred pairs of a new type of combat boots were tested for durability in Vietnam and, on the average, lasted two months.

{ (*i*) Homes in Karachi city having telephones and T.V.,
(*ii*) An infinite number of tosses of a coin,
(*iii*) Total production of a new type of combat boots during a particular period. }

(*c*) In each of the following situations, determine whether the sampling is done from a finite population or an infinite population and then define the population.

(*i*) A coin is tossed 20 times and 12 heads are recorded.

(*ii*) Ten employees of large manufacturing company are selected as representatives of labour to serve as a labour management committee.

(*iii*) A sample of bulbs is selected periodically to determine the number of defective bulbs produced by a production unit.

(*iv*) A coin is weighed 15 times to estimate its true weight.

{ (*i*) Infinite—population is all the potential tosses of the coin,
(*ii*) Finite—population is all the employees of the company,
(*iii*) Infinite—population is all the bulbs produced by the production unit,
(*iv*) Infinite—population is all the potential weights of the coin.}

2. (*a*) What is meant by sampling? Describe the advantages of sampling over complete enumeration.

(*b*) For each of the following reasons, give an example of a situation for which a census would be less desirable than a sample. In each case, explain why this is so,

(*i*) Economy (*ii*) Timeliness
(*iii*) Size of population (*iv*) Inaccessibility
(*v*) Accuracy (*vi*) Destructive observations?

(*c*) Distinguish between the following:

(*i*) Population and sample,

(*ii*) Sampling with and without replacement.

3. (*a*) Distinguish between probability and non-probability sampling, giving examples. Describe the advantages of using a probability sample.

(*b*) What do you understand by a simple random sample? By taking some artificial example, explain the method of drawing a simple random sample.

(*c*) Distinguish between the following:

(*i*) Random sampling and simple random sampling.

(*ii*) Simple random sampling and stratified random sampling.

(*iii*) Sampling and non-sampling errors.

4. (*a*) Explain how would you select a random sample of 10 households from a list of 250 households, by using a table of random digits.

(*b*) A poll is to be conducted to determine the voting preference of the voters in a certain city. Design a sampling plan such that the sample would be representative of the population of all the voters.

(*c*) In a certain locality, there are 300 households. We wish to select a sample of 50 households. How would you select this sample using Random Numbers?

5. (*a*) What is the difference between precision and accuracy of a result? Explain with some examples.

(*b*) What are two broad categories of errors in data collected by sample surveys? What are the methods for reducing sampling error?

◆

11.9 SIMPLE RANDOM SAMPLING AND SAMPLING DISTRIBUTIONS

As we have already mentioned that a random sample must be chosen in such a way that it is representative of the population about which we want to make inferences. A random sample of observations can be chosen in either of the two ways: with replacement or without replacement.

11.10 SAMPLING DISTRIBUTION OF A STATISTIC

11.10.1 Parameters. The numerical quantities, that describe probability distributions, are called *parameters*. Parameters are fixed constants that characterise a population.

Parameters are usually denoted by Greek letters. Thus, π (the probability of success in a binomial experiment), and μ and σ (the mean and standard deviation of a normal distribution) are examples of parameters.

Let $x_1, x_2, \cdots, x_N$ be the N elements of a population. A population value summarizes the values of some characteristic (or characteristics) for all N units of an entire population. It describes some feature of the distribution of the random variable (or variables) in the defined population. Let x_j, $j = 1, 2, \cdots, N$, be the observed value of some random variable X for the j-th element in the population, then some of the examples of the population parameters are:

Population total: $$\tau = \sum_{j=1}^{N} \sum x_j$$

Population mean: $$\mu = \frac{\sum_{j=1}^{N} x_j}{N}$$

Population variance: $$\sigma^2 = \frac{\sum_{j=1}^{N} (x_j - \mu)^2}{N}$$

Population proportion: $$\pi = \frac{No.of\ elements\ with\ attribute\ A}{Population\ size} = \frac{k}{N}$$

11.10.2 Statistic. A *statistic* is a function, of the observations of a random sample, which does not contain any unknown parameter.

We know that a number of simple random samples can be drawn from the same population and each sample gives a different value of the statistic that is used as an estimator of the population parameter. The sample statistic is a random variable having its own probability distribution. We intend to use a statistic to make inferences about the distribution of the population. A statistic is usually denoted by a small Latin letter ($\bar{x}$, s, r) to represent its value obtained from an actually observed sample. A statistic is denoted by a capital Latin letter ($\bar{X}$, S, R) to represent its random nature.

Let $x_1, x_2, \cdots, x_n$ be the observed values of a random sample $X_1, X_2, \cdots, X_n$ of size n from a given population of N items. A sample value is an estimate calculated from the n

elements in the sample. Let x_i, $i = 1, 2, \cdots, n$ be the i-th element in the sample, then the observed values of some of the sample statistics are

Sample total: $\sum_{i=1}^{n} x_i$

Sample mean: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Sample variance: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$, $\hat{s}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

where $n s^2 = (n-1)\hat{s}^2$

Sample proportion: $p = \dfrac{\text{No. of elements with attribute } A}{\text{Sample size}} = \dfrac{x}{n}$

11.10.3 Sampling Distribution of a Statistic. The *sampling distribution of statistic* is the probability distribution of the statistic obtained from all possible samples of some specified size that can be drawn from a given population.

11.10.4 Standard Error of a Statistic. The standard deviation of the sampling distribution of a statistic is called the *standard error of the statistic*.

11.11 SAMPLING DISTRIBUTIONS FROM GENERAL POPULATIONS

We will now look at the most common situations where the Central Limit Theorem is used to specify approximate probability distributions for sample statistics where sampling is done from general (non-normal) populations.

The particular sampling distributions we are interested in are those for: (*i*) the mean, (*ii*) the difference between two means, (*iii*) the proportion of successes, (*iv*) the difference between two proportions.

11.12 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN, $\bar{X}$

The *sampling distribution of the sample mean* $\bar{X}$ is the probability distribution of the means of all possible simple random samples of n observations that can be drawn from a given population with mean μ and variance σ^2.

11.12.1 Standard Error of $\bar{X}$. The standard deviation of the sampling distribution of the sample mean $\bar{X}$, denoted by $\sigma_{\bar{X}}$, is called the *standard error of* $\bar{X}$.

To discuss the relationships between the population and the sampling distribution of the sample mean, the following symbols will be used.

N = Population size	n = Sample size
μ = Population mean	$\mu_{\bar{X}}$ = Mean of the distribution of $\bar{X}$
σ^2 = Population variance	$\sigma^2_{\bar{X}}$ = Variance of the distribution of $\bar{X}$
σ = Population standard deviation	$\sigma_{\bar{X}}$ = Standard error of the distribution of $\bar{X}$

11.12.2 Properties of the Sampling Distribution of $\bar{X}$. The properties of the sampling distribution of the sample mean are given by the following theorems:

***Theorem* 11.2** *The mean of the sampling distribution of $\bar{X}$, denoted by $\mu_{\bar{X}}$, is equal to the mean of the sampled population, i. e.,*

$$\mu_{\bar{X}} = E(\bar{X}) = \mu$$

This theorem holds regardless of the sample size n or whether sampling is conducted with or without replacement.

***Theorem* 11.3** *The variance of the sampling distribution of $\bar{X}$ is equal to the variance of the sampled population divided by the sample size, i. e.,*

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2}{n}$$

where $\bar{X}$ is the mean of a random sample of size n from an infinite population (or sampling with replacement) with mean μ and finite variance σ^2.

The standard error of $\bar{X}$ then becomes

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

However, if the value of σ is unknown, it is replaced by the sample standard deviation $\hat{S}$, the estimate of the standard error of $\bar{X}$ then becomes

$$S_{\bar{X}} = \hat{\sigma}_{\bar{X}} = \frac{\hat{S}}{\sqrt{n}}$$

where, $\hat{S} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$

***Theorem* 11.4** *The variance of the sampling distribution of $\bar{X}$ is*

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

*where $\bar{X}$ is the mean of a random sample of size n drawn **without replacement** from a finite population of size N with mean μ and variance σ^2. The factor $(N-n)/(N-1)$ is usually called as finite population correction ($f p c$) for variance.*

The standard error of $\bar{X}$ then becomes

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

However, if the value of σ is unknown, it is replaced by the sample standard deviation $\hat{S}$, the estimate of the standard error of $\bar{X}$ then becomes

$$S_{\bar{X}} = \hat{\sigma}_{\bar{X}} = \frac{\hat{S}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

where, $\hat{S} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$

***Theorem* 11.5** *If $\bar{X}$ is the mean of the random sample of size n drawn from a normal population with mean μ and variance σ^2 (known), the sampling distribution of $\bar{X}$ is a normal distribution with mean μ and variance σ^2/n regardless of the size of the sample (including sample size 1). The distribution of the standardized sampling errors*

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

will be standard normal distribution.

***Theorem* 11.6 (Central Limit Theorem).** *For a large sample size, the mean $\bar{X}$ of a random sample from a population with mean μ and finite variance σ^2 has a sampling distribution that is approximately normal with mean μ and variance σ^2/n regardless of the probability distribution (shape) of the sampled population. The lager the sample size, the better will be the normal approximation to the sampling distribution of $\bar{X}$. The distribution of the standardized sampling errors*

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

will approach the standard normal distribution as n tends to infinity.

***Example* 11.1** *A population consists of four children with ages 2, 4, 6 and 8. Take all possible simple random samples of size 2 with replacement. If X is the age of a child, find,*

(i) the theoretical sampling distribution of $\bar{X}$, the mean age of two children in a sample;

(ii) the mean, variance and standard error of $\bar{X}$;

(iii) the mean, variance and standard deviation of the population.

Verify the results

(i) $\mu_{\bar{X}} = \mu$ *(ii)* $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ *(iii)* $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Solution. Population: 2, 4, 6, 8; Population size: $N = 4$; Sample size: $n = 2$

Number of possible samples $= N \times N = 4 \times 4 = 16$

All possible samples that can be drawn with replacement from our population, and their means are shown in the following tree diagram.

First draw	Second draw	Sample values x_i	Sample sum $\sum x_i$	$\bar{x} = \frac{\sum x_i}{n}$
2	2	2, 2	4	2
	4	2, 4	6	3
	6	2, 6	8	4
	8	2, 8	10	5
4	2	4, 2	6	3
	4	4, 4	8	4
	6	4, 6	10	5
	8	4, 8	12	6
6	2	6, 2	8	4
	4	6, 4	10	5
	6	6, 6	12	6
	8	6, 8	14	7
8	2	8, 2	10	5
	4	8, 4	12	6
	6	8, 6	14	7
	8	8, 8	16	8

Fig. 11.1 A tree diagram showing all possible samples of size 2 drawn with replacement from a population of the 4 equiprobable values 2, 4, 6, 8

The sampling distribution of sample mean $\bar{X}$, its mean, variance and standard error are

Value of $\bar{X}$	Number of occurrences	Probability		
$\bar{x}$	f	$p(\bar{x}) = f/\sum f$	$\bar{x}\, p(\bar{x})$	$\bar{x}^2\, p(\bar{x})$
2	1	1/16	2/16	4/16
3	2	2/16	6/16	18/16
4	3	3/16	12/16	48/16
5	4	4/16	20/16	100/16
6	3	3/16	18/16	108/16
7	2	2/16	14/16	98/16
8	1	1/16	8/16	64/16
Sums	$\sum f = 16$	1	80/16	440/16

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{x}\, p(\bar{x}) = \frac{80}{16} = 5$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \sum \bar{x}^2\, p(\bar{x}) - \mu^2_{\bar{X}} = \frac{440}{16} - (5)^2 = 2.5$$

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \sqrt{2.5} = 1.58$$

The mean, variance and standard deviation of the population

x_j	2	4	6	8	$\Sigma x_j = 20$
x_j^2	4	16	36	64	$\Sigma x_j^2 = 120$

$$\mu = \frac{\Sigma x_j}{N} = \frac{20}{4} = 5$$

$$\sigma^2 = \frac{\Sigma x_j^2}{N} - \mu^2 = \frac{120}{4} - (5)^2 = 5$$

$$\sigma = \sqrt{5} = 2.236$$

We are to verify that

(*i*) $\mu_{\bar{X}} = \mu$ $\quad 5 = 5$

(*ii*) $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ $\quad 2.5 = \frac{5}{2}$ $\quad 2.5 = 2.5$

(*iii*) $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ $\quad 1.58 = \frac{2.236}{\sqrt{2}}$ $\quad 1.58 = 1.58$

***Example* 11.2** *A population consists of values* 3, 6, *and* 9. *Take all possible simple random samples of size* 3 *with replacement. Form the sampling distribution of sample mean* $\bar{X}$. *Hence state and verify the relationship between*

(*i*) *the mean of* $\bar{X}$ *and the population mean,*

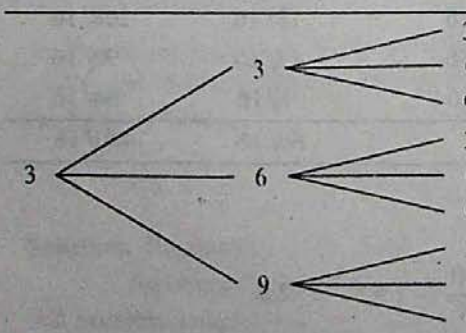
(*ii*) *the variance of* $\bar{X}$ *and the population variance,*

(*iii*) *the standard error of* $\bar{X}$ *and the population standard deviation.*

Solution. Population: 3, 6, 9; Population size: $N = 3$; Sample size: $n = 3$

Number of possible samples $= N \times N \times N = 3 \times 3 \times 3 = 27$

All possible samples that can be drawn with replacement from our population, and the sample means are shown in the following tree diagram.



First draw	Second draw	Third draw	Sample values x_i	Σx_i	$\bar{x} = \frac{\Sigma x_i}{n}$
3	3	3	3, 3, 3	9	3
		6	3, 3, 6	12	4
		9	3, 3, 9	15	5
	6	3	3, 6, 3	12	4
		6	3, 6, 6	15	5
		9	3, 6, 9	18	6
	9	3	3, 9, 3	15	5
		6	3, 9, 6	18	6
		9	3, 9, 9	21	7

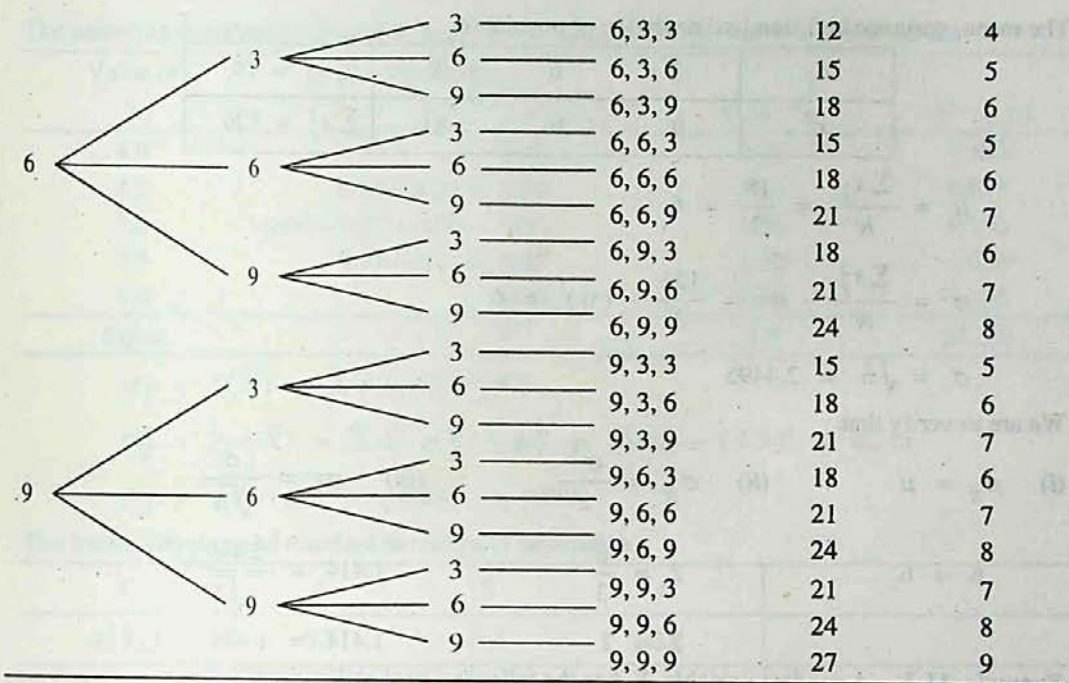


Fig. 11.2 A tree diagram showing all possible samples of size 3 drawn with replacement from a population of 3 equiprobable values 3, 6, 9.

The sampling distribution of sample mean $\bar{X}$, its mean, variance and standard error are

Value of $\bar{x}$	Number of occurrences f	Probability $p(\bar{x}) = f/\sum f$	$\bar{x}\ p(\bar{x})$	$\bar{x}^2\ p(\bar{x})$
3	1	1/27	3/27	9/27
4	3	3/27	12/27	48/27
5	6	6/27	30/27	150/27
6	7	7/27	42/27	252/27
7	6	6/27	42/27	294/27
8	3	3/27	24/27	192/27
9	1	1/27	9/27	81/27
Sum	$\sum f = 27$	1	162/27	1026/27

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{x}\ p(\bar{x}) = \frac{162}{27} = 6$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \sum \bar{x}^2\ p(\bar{x}) - \mu^2_{\bar{X}} = \frac{1026}{27} - (6)^2 = 2$$

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \sqrt{2} = 1.414$$

The mean, variance and standard deviation of population

x_j	3	6	9	$\sum x_j = 18$
x_j^2	9	36	81	$\sum x_j^2 = 126$

$$\mu = \frac{\sum x_j}{N} = \frac{18}{3} = 6$$

$$\sigma^2 = \frac{\sum x_j^2}{N} - \mu^2 = \frac{126}{3} - (6)^2 = 6$$

$$\sigma = \sqrt{6} = 2.4495$$

We are to verify that

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ (*iii*) $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

$6 = 6$ $2 = \frac{6}{3}$ $1.414 = \frac{2.4495}{\sqrt{3}}$

$2 = 2$ $1.414 = 1.414$

***Example* 11.3** *A random variable X has the following probability distribution.*

x_j	4	5	6
$p(x_j)$	0.3	0.5	0.2

If a sample of size 2 is taken with replacement, obtain the sampling distribution of $\bar{X}$. Determine the mean and variance of the sampling distribution. Find the mean and variance of the population. Discuss the results.

Solution. We have an infinite population. Since the sample is drawn at random with replacement from the *infinite population*, the sample values are independent. Thus the distribution of possible samples of size $n = 2$ drawn with replacement is

Sample values x_i	Sample total $\sum x_i$	Sample mean $\bar{x} = \frac{\sum x_i}{n}$	Probability $p(\bar{x})$
4, 4	8	4.0	(0.3)(0.3) = 0.09
4, 5	9	4.5	(0.3)(0.5) = 0.15
4, 6	10	5.0	(0.3)(0.2) = 0.06
5, 4	9	4.5	(0.5)(0.3) = 0.15
5, 5	10	5.0	(0.5)(0.5) = 0.25
5, 6	11	5.5	(0.5)(0.2) = 0.10
6, 4	10	5.0	(0.2)(0.3) = 0.06
6, 5	11	5.5	(0.2)(0.5) = 0.10
6, 6	12	6.0	(0.2)(0.2) = 0.04
Sum			1

The sampling distribution of sample mean $\bar{X}$, its mean, variance and standard error are

Value of $\bar{x}$	Probability $p(\bar{x})$	$\bar{x}\ p(\bar{x})$	$\bar{x}^2\ p(\bar{x})$
4.0	0.09	0.36	1.440
4.5	0.15 + 0.15 = 0.30	1.35	6.075
5.0	0.06 + 0.25 + 0.06 = 0.37	1.85	9.250
5.5	0.10 + 0.10 = 0.20	1.10	6.050
6.0	0.04	0.24	1.440
Sums	1	4.9	24.255

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{x}\, p(\bar{x}) = 4.9$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \sum \bar{x}^2\, p(\bar{x}) - \mu^2_{\bar{X}} = 24.255 - (4.9)^2 = 0.245$$

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \sqrt{0.245} = 0.495$$

The mean, variance and standard deviation of population

x_j	4	5	6	
$p(x_j)$	0.3	0.5	0.2	
$x_j\ p(x_j)$	1.2	2.5	1.2	$\sum x_j\ p(x_j) = 4.9$
$x_j^2\ p(x_j)$	4.8	12.5	7.2	$\sum x_j^2\ p(x_j) = 24.5$

$$\mu = E(X) = \sum x_j\ p(x_j) = 4.9$$

$$\sigma^2 = Var(X) = \sum x_j^2\ p(x_j) - \mu^2 = 24.5 - (4.9)^2 = 0.49$$

$$\sigma = \sqrt{0.49} = 0.7$$

We are to verify that:

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n}$ (*iii*) $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

$4.9 = 4.9$ $0.245 = \dfrac{0.49}{2} = 0.245$ $0.495 = \dfrac{0.7}{\sqrt{2}} = 0.495$

***Example* 11.4** *A population consists of value* 3, 5, 7 *and* 9. *Take all possible simple random samples of size* 2 *without replacement. Form the sampling distribution of sample mean* $\bar{X}$. *Find the mean, variance and standard error of* $\bar{X}$. *Find the mean, variance and standard deviation of the population. Verify that*:

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$ (*iii*) $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

Solution. Population: 3, 5, 7, 9; Population size: $N = 4$; Sample size: $n = 2$

Number of possible samples $= N(N-1) = 4(4-1) = 12$

All possible samples that can be drawn without replacement from our population, and their means are shown in the following tree diagram.

First draw	Second draw	Sample values x_i	$\sum x_i$	$\bar{x} = \frac{\sum x_i}{n}$
3	5	3, 5	8	4
	7	3, 7	10	5
	9	3, 9	12	6
5	3	5, 3	8	4
	7	5, 7	12	6
	9	5, 9	14	7
7	3	7, 3	10	5
	5	7, 5	12	6
	9	7, 9	16	8
9	3	9, 3	12	6
	5	9, 5	14	7
	7	9, 7	16	8

Fig. 11.3 A tree diagram showing all possible samples of size 2 drawn without replacement from a population of 4 equiprobable values 3, 5, 7, 9.

The sampling distribution of sample mean $\bar{X}$, its mean, variance and standard error are

Value of $\bar{X}$ $\bar{x}$	Number of occurrences f	Probability $p(\bar{x}) = f/\sum f$	$\bar{x}\ p(\bar{x})$	$\bar{x}^2\ p(\bar{x})$
4	2	2/12	8/12	32/12
5	2	2/12	10/12	50/12
6	4	4/12	24/12	144/12
7	2	2/12	14/12	98/12
8	2	2/12	16/12	128/12
Sum	$\sum f = 12$	1	72/12	452/12

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{x}\ p(\bar{x}) = \frac{72}{12} = 6$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \sum \bar{x}^2\ p(\bar{x}) - \mu^2_{\bar{X}} = \frac{452}{12} - (6)^2 = 1.667$$

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \sqrt{1.667} = 1.291$$

The mean, variance and standard deviation of population

x_j	3	5	7	9	$\sum x_j = 24$
x_j^2	9	25	49	81	$\sum x_j^2 = 164$

$$\mu = \frac{\sum x_j}{N} = \frac{24}{4} = 6$$

$$\sigma^2 = \frac{\sum x_j^2}{N} - \mu^2 = \frac{164}{4} - (6)^2 = 5$$

$$\sigma = \sqrt{5} = 2.236$$

We are to verify that

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ (*iii*) $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

$6 = 6$ $1.667 = \frac{5}{2}\left(\frac{4-2}{4-1}\right)$ $1.291 = \frac{2.236}{\sqrt{2}}\sqrt{\frac{4-2}{4-1}}$

$1.667 = 1.667$ $1.291 = 1.291$

***Example* 11.5** *A population consists of values* 0, 3, 6 *and* 9. *Take all possible simple random samples of size* 3 *without replacement. Form the sampling distribution of sample mean* $\bar{X}$. *Hence state and verify the relationship between*

(*i*) *the mean of* $\bar{X}$ *and the population mean,*

(*ii*) *the variance of* $\bar{X}$ *and the population variance,*

(*iii*) *the standard error of* $\bar{X}$ *and the population standard deviation.*

Solution. Population: 0, 3, 6, 9; Population size: $N = 4$; Sample size: $n = 3$

Number of possible samples $= N(N-1)(N-2) = 4(4-1)(4-2) = 24$

All possible samples that can be drawn without replacement from our population, and the sample means are shown in the following tree diagram.

First draw	Second draw	Third draw	Sample values x_i	$\sum x_i$	$\bar{x} = \frac{\sum x_i}{n}$
0	3	6	0, 3, 6	9	3
		9	0, 3, 9	12	4
	6	3	0, 6, 3	9	3
		9	0, 6, 9	15	5
	9	3	0, 9, 3	12	4
		6	0, 9, 6	15	5

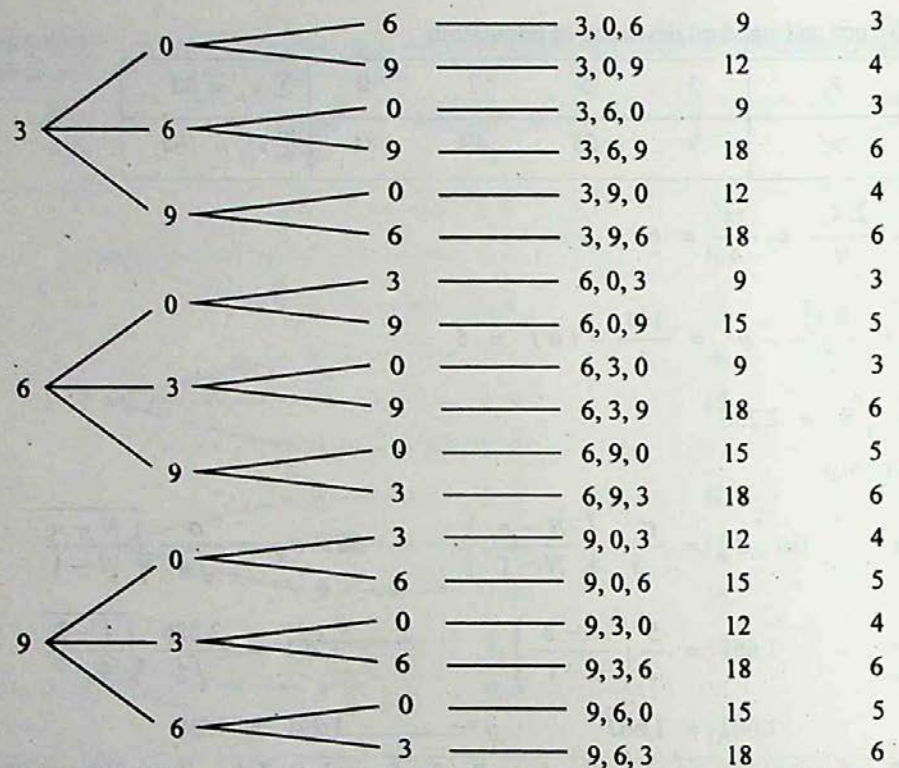


Fig. 11.4 A tree diagram showing all possible samples of size 3 drawn without replacement from a population of 4 equiprobable values 0, 3, 6, 9.

The sampling distribution of sample mean $\overline{X}$, its mean, variance and standard error are

Value of $\overline{X}$	Number of occurrences	Probability		
$\overline{x}$	f	$p(\overline{x}) = f/\Sigma f$	$\overline{x}\ p(\overline{x})$	$\overline{x}^2\ p(\overline{x})$
3	6	6/24	18/24	54/24
4	6	6/24	24/24	96/24
5	6	6/24	30/24	150/24
6	6	6/24	36/24	216/24
Sums	$\Sigma f = 24$	1	108/24	516/24

$$\mu_{\overline{X}} = E(\overline{X}) = \Sigma \overline{x}\ p(\overline{x}) = \frac{108}{24} = 4.5$$

$$\sigma^2_{\overline{X}} = Var(\overline{X}) = \Sigma \overline{x}^2\ p(\overline{x}) - \mu^2_{\overline{X}} = \frac{516}{24} - (4.5)^2 = 1.25$$

$$\sigma_{\overline{X}} = \sqrt{Var(\overline{X})} = \sqrt{1.25} = 1.118$$

The mean, variance and standard deviation of population

x_j	0	3	6	9	$\Sigma x_j = 18$
x_j^2	0	9	36	81	$\Sigma x_j^2 = 126$

$$\mu = \frac{\Sigma x_j}{N} = \frac{18}{4} = 4.5$$

$$\sigma^2 = \frac{\Sigma x_j^2}{N} - \mu^2 = \frac{126}{4} - (4.5)^2 = 11.25$$

$$\sigma = \sqrt{11.25} = 3.3541$$

We are to verify that

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ (*iii*) $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

$4.5 = 4.5$ $\quad 1.25 = \frac{11.25}{3}\left(\frac{4-3}{4-1}\right)$ $\quad 1.118 = \frac{3.3541}{\sqrt{3}}\sqrt{\frac{4-3}{4-1}}$

$1.25 = 1.25$ $\quad 1.118 = 1.118$

***Example* 11.6** *A random variable X has the following probability distribution.*

x_j	3	4	5
$p(x_j)$	0.2	0.4	0.4

If a simple random sample of 3 numbers is taken without replacement, obtain a sampling distribution of the sample mean $\bar{X}$. Find the mean, variance and standard error of $\bar{X}$.

Solution. We have an infinite population. The actual sampling distribution of $\bar{X}$, the sample mean of three numbers taken without replacement, is impracticable because the population is infinite. Since the sample is drawn at random without replacement from the infinite population, the sample values become independent. Then the actual sampling distribution of $\bar{X}$, the sample mean of three numbers taken without replacement, is impossible but it virtually becomes the sampling distribution of $\bar{X}$ the sample mean of three numbers taken with replacement.

The population size N is infinite and $n = 3$. Then the finite population correction

$$\frac{N-n}{N-1} \to 1 \text{ as } N \to \infty \quad \text{and} \quad \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \to \frac{\sigma}{\sqrt{n}}$$

The mean, variance and standard deviation of population

x_j	3	4	5	Sum
$p(x_j)$	0.2	0.4	0.4	$\Sigma p(x_j) = 1$
$x_j\, p(x_j)$	0.6	1.6	2.0	$\Sigma x_j\, p(x_j) = 4.2$
$x_j^2\, p(x_j)$	1.8	6.4	10.0	$\Sigma x_j^2\, p(x_j) = 18.2$

$$\mu = E(X) = \sum x p(x) = 4.2$$

$$\sigma^2 = Var(X) = \sum x^2 p(x) - \mu^2 = 18.2 - (4.2)^2 = 0.56$$

The mean, variance and standard error of $\bar{X}$

$$\mu_{\bar{X}} = \mu = 4.2$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{0.56}{3} = 0.187$$

$$\sigma_{\bar{X}} = \sqrt{0.187} = 0.432$$

***Example* 11.7** *The weights of* 1000 *students of a college are normally distributed with mean* 68.5 *kg and standard deviation* 2.7 *kg. If a simple random sample of* 25 *students is obtained from this population, find the expected mean and standard deviation of the sampling distribution of means if sampling were done* (*i*) *with replacement and* (*ii*) *without replacement.*

Solution. We have

Population mean: $\mu = 68.5$, Population standard deviation: $\sigma = 2.7$

Population size: $N = 1000$, Sample size: $n = 25$

(*i*) ***Sampling with replacement*:**

$$\mu_{\bar{X}} = \mu = 68.5 \text{ kg.}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.7}{\sqrt{25}} = 0.54 \text{ kg}$$

(*ii*) ***Sampling without replacement*:**

$$\mu_{\bar{X}} = \mu = 68.5 \text{ kg.}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{2.7}{\sqrt{25}} \sqrt{\frac{1000-25}{1000-1}} = 0.53 \text{ kg}$$

***Example* 11.8** *Given the population* 1, 1, 1, 3, 4, 5, 6, 6, 6, *and* 7.

(*a*) *Find the mean and standard deviation for the sampling distribution of mean for a sample of size* 36 *selected at random with replacement.*

(*b*) *Find the mean and standard deviation for the sampling distribution of mean for a sample of size* 4 *selected at random without replacement.*

Solution. The mean and standard deviation of the population are:

x_j	1	1	1	3	4	5	6	6	6	7	$\sum x_j = 40$
x_j^2	1	1	1	9	16	25	36	36	36	49	$\sum x_j^2 = 210$

$$\mu = \frac{\sum x_j}{N} = \frac{40}{10} = 4$$

$$\sigma = \sqrt{\frac{\sum x_j^2}{N} - \mu^2} = \sqrt{\frac{210}{10} - (4)^2} = 2.236$$

(a) ***Sampling With Replacement.*** We have, sample size $n = 36$. The mean and standard error of $\overline{X}$ are:

$$\mu_{\overline{X}} = \mu = 4$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{\sqrt{36}} = 0.373$$

(b) ***Sampling Without Replacement.*** We have, sample size $n = 4$. The mean and standard error of $\overline{X}$ are:

$$\mu_{\overline{X}} = \mu = 4$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2.236}{\sqrt{4}}\sqrt{\frac{10-4}{10-1}} = 0.913$$

Exercise 11.2

1. **(a)** How do you define a population and a sample? Differentiate between parameter and statistic. Why a parameter is said to be a constant and statistic a variable?

(b) A labour union has 1000 members. A random sample of 50 members of the union gave an average age of 40 years. The average age of the members of the labour union was, therefore, estimated to be 40 years. A complete enumeration of all the members indicated that the true mean age was 43 years. Answer the following:

(*i*) Which figure is a parameter?

(*ii*) Which figure is a statistic?

{ (*i*) Population size $N = 1000$, and population mean age $\mu = 43$ years;
(*ii*) Sample size n, and sample mean $\overline{x} = 40$ years.}

2. **(a)** What is meant by a sampling distribution and a standard error? Describe the properties of the sampling distribution of sample mean.

(b) What is meant by standard error and what are its practical uses?

(c) What is the finite population correction factor? When is it appropriately used in sampling applications and when can it, without too great an undesirable consequence, be ignored?

3. **(a)** A finite population consists of the numbers 2, 4, 6, 8, 10 and 12. Calculate the sample means for all possible random samples of size $n = 2$, that can be drawn from this population, with replacement. Assuming the 36 possible samples equally likely, make the sampling distribution of sample means and find the mean and variance of this distribution. Calculate mean and variance of the population and verify that

(*i*) $\mu_{\overline{X}} = \mu$ (*ii*) $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$

{ $\mu = 7$, $\sigma^2 = 11.667$, $\mu_{\overline{X}} = 7$, $\sigma^2_{\overline{X}} = 5.833$ }

(*b*) A finite population consists of the numbers 2, 4, 6, 6, 8 and 10. Calculate the sample means for all possible random samples of size $n = 2$, that can be drawn from this population, with replacement. Assuming the 36 possible samples equally likely, form the sampling distribution of sample means and find the mean and variance of this distribution. Calculate mean and variance of the population and verify that

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

$\{\mu = 6,\ \mu_{\bar{X}} = 6,\ \sigma = 2.582,\ \sigma_{\bar{X}} = 1.826\}$

(*c*) Draw all possible samples of size $n = 3$ with replacement from the population 3, 6, 9 and 12. Assuming the 64 possible samples equally likely, form a sampling distribution of the sample means. Hence state and verify the relation between

(*i*) the mean of the sampling distribution of the sample mean and the population mean;

(*ii*) the variance of the sampling distribution of the sample mean and the population variance.

$\{\mu = 7.5,\ \mu_{\bar{X}} = 7.5,\ \mu_{\bar{X}} = \mu;\ \sigma^2 = 11.25,\ \sigma^2_{\bar{X}} = 3.75,\ \sigma^2_{\bar{X}} = \sigma^2/n\}$

4. (*a*) A finite population consists of the numbers 2, 4, 6, 6, 8 and 10. Calculate the sample means for all possible random samples of size $n = 2$, that can be drawn from this population, without replacement. Assuming the 30 possible samples equally likely, make the sampling distribution of sample mean. Find the mean and variance of this distribution. Calculate mean and variance of the population and verify that

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

$\{\mu = 6,\ \mu_{\bar{X}} = 6,\ \sigma = 2.582,\ \sigma_{\bar{X}} = 1.633\}$

(*b*) A finite population consists of the values 6, 6, 9, 15 and 18. Calculate the sample means for all possible random samples of size $n = 3$, that can be drawn from this population, without replacement. Assuming the 60 possible samples equally likely, make the sampling distribution of sample mean and find the mean and variance of this distribution. Calculate mean and variance of the population and show that

(*i*) $\mu_{\bar{X}} = \mu$ (*ii*) $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$

$(\mu = 10.8,\quad \sigma^2 = 23.76,\quad \mu_{\bar{X}} = 10.8,\quad \sigma^2_{\bar{X}} = 3.96)$

(*c*) Find the mean μ and variance σ^2 of the finite population 1, 4, 7 and 8. Take all possible samples of size 2, that can be drawn at random without replacement from this population. Assuming the 12 possible samples equally likely, make the sampling distribution of sample mean and find the mean and variance of this distribution. Verify that

(*i*) $E(\bar{X}) = \mu$ (*ii*) $Var(\bar{X}) = \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$

where $\overline{X}$ is the random variable 'the sample mean', N is the population size and n is the sample size. What happens as $N \to \infty$?

$\{ \mu = 5,\ \sigma^2 = 7.5,\ E(\overline{X}) = 5,\ Var(\overline{X}) = 2.5,\ \frac{N-n}{N-1} \to 1 \}$

5. (*a*) In an infinite population $\mu = 50$ and $\sigma^2 = 250$, find the mean and variance for the distribution of $\overline{X}$ if :

(*i*) $n = 25$, (*ii*) $n = 100$, (*iii*) $n = 1250$

$\{$ (*i*) $\mu_{\overline{X}} = 50,\ \sigma^2_{\overline{X}} = 10$ (*ii*) $\mu_{\overline{X}} = 50,\ \sigma^2_{\overline{X}} = 2.5$ (*iii*) $\mu_{\overline{X}} = 50,\ \sigma^2_{\overline{X}} = 0.2$ $\}$

(*b*) A large number of samples of size 50 were selected at random from a normal population with mean μ and variance σ^2. The mean and standard error of the sampling distribution of the sample mean were obtained 2500 and 4 respectively. Find the mean and variance of the population.

(2500, 800)

6. (*a*) If the size of the simple random sample from an infinite population is 55, the variance of sample mean is 27, what must be the standard error of sample mean if $n = 165$?

($\sigma_{\overline{X}} = 3$)

(*b*) If the size of the simple random sample from an infinite population is 36 and the standard error of the mean is 2, what must the size of the sample become if the standard error is to be reduced to 1.2?

($n = 100$)

7. (*a*) The random variable X has the following probability distribution:

x_j	4	5	6	7
$p(x_j)$	0.2	0.4	0.3	0.1

Find the mean $\mu_{\overline{X}}$, variance $\sigma^2_{\overline{X}}$ and standard error $\sigma_{\overline{X}}$ of the mean $\overline{X}$ for a random sample of size 36.

($\mu_{\overline{X}} = 5.3,\ \sigma^2_{\overline{X}} = 0.0225,\ \sigma_{\overline{X}} = 0.15$)

(*b*) A random sample of 36 cases is drawn from a negatively skewed probability distribution with a mean of 2 and a standard deviation of 3. Find the mean and standard error of the of the sampling distribution of $\overline{X}$.

($\mu_{\overline{X}} = 2,\ \sigma_{\overline{X}} = 0.5$)

(*c*) A random sample of 100 is taken from a population with mean 30 and standard deviation 5. The probability distribution of the parent population is unknown, find the mean and standard error of the of the sampling distribution of $\overline{X}$.

($\mu_{\overline{X}} = 30,\ \sigma_{\overline{X}} = 0.5$)

◆

11.13 SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS, $\bar{X}_1 - \bar{X}_2$

The *sampling distribution of the difference between two sample means* $\bar{X}_1 - \bar{X}_2$ is the probability distribution of all possible differences between means $\bar{X}_1$ and $\bar{X}_2$ obtained from all possible independent simple random samples of n_1 and n_2 observations that can be drawn from two given populations with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively.

Often we wish to compare the means of two random variables. The comparison is made on the basis of two independent random samples drawn from given populations.

Suppose that two independent random samples of sizes n_1 and n_2 are drawn from populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. Let $\bar{X}_1$ be the mean of sample of size n_1 from the population with mean μ_1 and variance σ_1^2, then $\bar{X}_1$ is a random variable that has its own probability distribution with mean μ_1 and variance σ_1^2/n_1. Let $\bar{X}_2$ be the mean of sample of size n_2 from the population with mean μ_2 and variance σ_2^2, then $\bar{X}_2$ is a random variable that has its own probability distribution with mean μ_2 and variance σ_2^2/n_2.

Then the differences $\bar{X}_1 - \bar{X}_2$ can be obtained from all possible pairs of $\bar{X}_1$ and $\bar{X}_2$. Consequently, the difference $\bar{X}_1 - \bar{X}_2$ between two sample means is a random variable that has its own probability distribution which is called the sampling distribution of the difference between two sample means.

11.13.1 Properties of the Sampling Distribution of the Difference between Two Sample Means. The properties of the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ between two sample means are given by the following theorems:

***Theorem* 11.7** *The mean of the sampling distribution of* $(\bar{X}_1 - \bar{X}_2)$, *denoted by* $\mu_{\bar{X}_1 - \bar{X}_2}$, *is equal to the difference between the population means, i. e.,*

$$\mu_{\bar{X}_1 - \bar{X}_2} = E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

This theorem holds regardless of the sample sizes n_1 and n_2 or whether sampling is done with or without replacement.

***Theorem* 11.8** *The variance of the sampling distribution of* $(\bar{X}_1 - \bar{X}_2)$, *denoted by* $\sigma^2_{\bar{X}_1 - \bar{X}_2}$, *is equal to sum of the variances of the sampled populations divided by the respective sample sizes, i. e.,*

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

where $\bar{X}_1$ *and* $\bar{X}_2$ *are means of two independent random samples of sizes* n_1 *and* n_2 *from infinite populations (or sampling with replacement) with means* μ_1 *and* μ_2 *and finite variances* σ_1^2 *and* σ_2^2 *respectively.*

The standard error of $(\bar{X}_1 - \bar{X}_2)$ then becomes

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{Var(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

However, if σ_1^2 and σ_2^2 are unknown, these are replaced by the sample variances $\hat{S}_1^2$ and $\hat{S}_2^2$, the estimate of the standard error of $(\bar{X}_1 - \bar{X}_2)$ then becomes

$$S_{\bar{X}_1 - \bar{X}_2} = \hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

where $\hat{S}_1^2 = \dfrac{\Sigma(X_{i1} - \bar{X}_1)^2}{n_1 - 1}$ and $\hat{S}_2^2 = \dfrac{\Sigma(X_{i2} - \bar{X}_2)^2}{n_2 - 1}$

***Theorem* 11.9** *The variance of the sampling distribution of* $(\bar{X}_1 - \bar{X}_2)$ *is*

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)$$

where $\bar{X}_1$ *and* $\bar{X}_2$ *are the means of random samples of sizes* n_1 *and* n_2 *drawn without replacement from finite populations of sizes* N_1 *and* N_2 *with means* μ_1 *and* μ_2 *and variances* σ_1^2 *and* σ_2^2, *respectively.*

The standard error of $(\bar{X}_1 - \bar{X}_2)$ then becomes

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{Var(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

***Theorem* 11.10** *If* $\bar{X}_1$ *and* $\bar{X}_2$ *are the means of random samples of* n_1 *and* n_2 *observations from two independent normal populations with means* μ_1, μ_2 *and variances* σ_1^2, σ_2^2 *respectively, then the sampling distribution of the difference between sample means* $\bar{X}_1 - \bar{X}_2$ *is normal with mean and variance*

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

That is, the distribution of the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1 - \bar{X}_2}}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is a standard normal distribution.

***Example* 11.9** *Let* $\overline{X}_1$ *represent the mean of a sample of size* $n_1 = 2$ *selected at random with replacement from a finite population consisting of values* 7 *and* 9. *Similarly, let* $\overline{X}_2$ *represent the mean of a sample of size* $n_2 = 3$ *selected at random with replacement from another finite population consisting of values* 3 *and* 6. *Form a sampling distribution of the random variable* $(\overline{X}_1 - \overline{X}_2)$. *Verify that*

(*i*) $\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$ (*ii*) $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

Solution. We have

***Population* I:** 7, 9; $N_1 = 2$; $n_1 = 2$

Number of possible samples $= N_1 \times N_1 = 2 \times 2 = 4$

Possible samples (7, 7), (7, 9), (9, 7), (9, 9)

Sample means $\overline{x}_1$ 7, 8, 8, 9

***Population* II:** 3, 6; $N_2 = 2$; $n_2 = 3$

Number of possible samples $= N_2 \times N_2 \times N_2 = 2 \times 2 \times 2 = 8$

Possible samples (3, 3, 3), (3, 3, 6), (3, 6, 3), (6, 3, 3), (3, 6, 6), (6, 3, 6), (6, 6, 3), (6, 6, 6)

Sample means $\overline{x}_2$ 3, 4, 4, 4, 5, 5, 5, 6

All possible differences between sample means $(\overline{X}_1 - \overline{X}_2)$ are

$\overline{x}_1$	$\overline{x}_2$							
	3	4	4	4	5	5	5	6
7	4	3	3	3	2	2	2	1
8	5	4	4	4	3	3	3	2
8	5	4	4	4	3	3	3	2
9	6	5	5	5	4	4	4	3

The sampling distribution of $\overline{X}_1 - \overline{X}_2$, its mean and variance are

Value of $\overline{X}_1 - \overline{X}_2$	Number of occurrences	Probability		
$\overline{x}_1 - \overline{x}_2$	f	$p(\overline{x}_1 - \overline{x}_2) = f/\Sigma f$	$(\overline{x}_1 - \overline{x}_2)\, p(\overline{x}_1 - \overline{x}_2)$	$(\overline{x}_1 - \overline{x}_2)^2 p(\overline{x}_1 - \overline{x}_2)$
1	1	1/32	1/32	1/32
2	5	5/32	10/32	20/32
3	10	10/32	30/32	90/32
4	10	10/32	40/32	160/32
5	5	5/32	25/32	125/32
6	1	1/32	6/32	36/32
Sums	$\Sigma f = 32$	1	112/32	432/32

$$\mu_{\bar{X}_1 - \bar{X}_2} = E(\bar{X}_1 - \bar{X}_2) = \sum (\bar{x}_1 - \bar{x}_2)\, p(\bar{x}_1 - \bar{x}_2) = \frac{112}{32} = 3.5$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = Var(\bar{X}_1 - \bar{X}_2) = \sum (\bar{x}_1 - \bar{x}_2)^2\, p(\bar{x}_1 - \bar{x}_2) - \mu^2_{\bar{X}_1 - \bar{X}_2}$$

$$= \frac{432}{32} - (3.5)^2 = 1.25$$

The mean and variance of population I are

x_1	7	9	$\sum x_1 = 16$
x_1^2	49	81	$\sum x_1^2 = 130$

$$\mu_1 = \frac{\sum x_1}{N_1} = \frac{16}{2} = 8$$

$$\sigma_1^2 = \frac{\sum x_1^2}{N_1} - \mu_1^2 = \frac{130}{2} - (8)^2 = 1$$

The mean and variance of population II are

x_2	3	6	$\sum x_2 = 9$
x_2^2	9	36	$\sum x_2^2 = 45$

$$\mu_2 = \frac{\sum x_2}{N_2} = \frac{9}{2} = 4.5$$

$$\sigma_2^2 = \frac{\sum x_2^2}{N_2} - \mu_2^2 = \frac{45}{2} - (4.5)^2 = 2.25$$

We are to verify that

(*i*). $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ (*ii*) $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$3.5 = 8 - 4.5$ $1.25 = \frac{1}{2} + \frac{2.25}{3}$

$3.5 = 3.5$ $1.25 = 1.25$

***Example* 11.10** *Two independent random samples of sizes* $n_1 = 30$ *and* $n_2 = 50$ *are taken from two populations having means* $\mu_1 = 78$ *and* $\mu_2 = 75$ *and variances* $\sigma_1^2 = 150$ *and* $\sigma_2^2 = 200$. *Let* $\bar{X}_1$ *be the mean of the first random sample and* $\bar{X}_2$ *be the mean of the second random sample. Find the mean and standard error of* $\bar{X}_1 - \bar{X}_2$.

Solution. We have $\mu_1 = 78$, $\sigma_1^2 = 150$, $n_1 = 30$

$\mu_2 = 75$, $\sigma_2^2 = 200$, $n_2 = 50$

Then mean and standard error of $\bar{X}_1 - \bar{X}_2$ are

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 78 - 75 = 3$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{150}{30} + \frac{200}{50}} = 3$$

Exercise 11.3

1. (*a*) What is meant by the sampling distribution of the difference between two sample means. Describe the properties of the sampling distribution of the differences between two sample means.

(*b*) Let $\bar{X}_1$ represent the mean of a sample of size $n_1 = 2$, selected with replacement from a finite population -2, 0, 2, and 4. Similarly, let $\bar{X}_2$ represent the mean of a sample of size $n_2 = 2$, selected with replacement from the population -1 and 1.

(*i*) Assuming that the 64 possible differences $\bar{X}_1 - \bar{X}_2$ are equally likely to occur, construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$.

(*ii*) Verify that $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

$\{\mu_1 = 1, \mu_2 = 0, \mu_{\bar{X}_1 - \bar{X}_2} = 1, \sigma_1^2 = 5, \sigma_2^2 = 1, \sigma^2_{\bar{X}_1 - \bar{X}_2} = 3\}$

2. (*a*) Let the variable $\bar{X}_1$ represent the mean of random samples of size $n_1 = 2$, with replacement drawn from the finite population 3, 4, 5. Similarly, let $\bar{X}_2$ represent the means of random samples of size $n_2 = 3$, with replacement, drawn from the population 0, 3. Assuming that the 72 possible differences $\bar{X}_1 - \bar{X}_2$ are equally likely to occur, construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$. Show that

(*i*) $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ (*ii*) $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

$\{\mu_1 = 4, \mu_2 = 1.5, \mu_{\bar{X}_1 - \bar{X}_2} = 2.5, \sigma_1^2 = 0.667, \sigma_2^2 = 2.25, \sigma^2_{\bar{X}_1 - \bar{X}_2} = 1.083\}$

(*b*) Let the variable $\bar{X}_1$ represent the means of random samples of size 2 without replacement, drawn from the finite population 5, 7, 9. Similarly, let $\bar{X}_2$ represent the means of random sample of size 2, without replacement from another finite population 4, 6, 8. Assuming that the 36 possible differences $\bar{X}_1 - \bar{X}_2$ are equally likely to occur, construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and verify that

(*i*) $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

(*ii*) $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\sigma_1^2}{n_1}\left(\dfrac{N_1 - n_1}{N_1 - 1}\right) + \dfrac{\sigma_2^2}{n_2}\left(\dfrac{N_2 - n_2}{N_2 - 1}\right)$

$\{\mu_1 = 7, \mu_2 = 6, \mu_{\bar{X}_1 - \bar{X}_2} = 1, \sigma_1^2 = 2.667, \sigma_2^2 = 2.667\ \sigma^2_{\bar{X}_1 - \bar{X}_2} = 1.333\}$

3. (*a*) The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 years, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 years. A random sample of size 36 tubes is selected from manufacturer A and its mean $\bar{X}_1$ is calculated. An other random sample of size 49 tubes is selected from manufacturer B and its mean $\bar{X}_2$ is calculated. Find the mean and standard error of the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$. $(\mu_{\bar{X}_1 - \bar{X}_2} = 0.5, \sigma_{\bar{X}_1 - \bar{X}_2} = 0.1886)$

(*b*) Random samples of each size 100 are drawn from two independent probability distributions and their means $\bar{X}_1$ and $\bar{X}_2$ computed. If the means and standard deviations of the two populations are $\mu_1 = 10$, $\sigma_1 = 2$, $\mu_2 = 8$, $\sigma_2 = 1$, find the mean and standard error of the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$. $(\mu_{\bar{X}_1 - \bar{X}_2} = 2, \sigma_{\bar{X}_1 - \bar{X}_2} = 0.2236)$

11.14 SAMPLING DISTRIBUTION OF SAMPLE PROPORTION, *P*

The *sampling distribution of sample proportion* P is the probability distribution of the proportions of successes obtained from all possible simple random samples of n observations that can be drawn from a Bernoulli population with proportion of successes π.

11.14.1 Population Proportion. The *population proportion* is defined as

$$\pi = \frac{\textit{No. of elements with attribute } A}{\textit{Population size}} = \frac{k}{N}$$

where k is the number of elements in the population of size N that possess a certain characteristic. In many applications of sampling the characteristic of interest in the population elements is qualitative with two possible outcomes. Quite often, however, we are interested not in the number of successes but rather in the proportion of successes.

11.14.2 Sample Statistics *X* and *P*. When the characteristic of interest is qualitative with two possible outcomes, a sample statistic of interest is the *number of occurrences* among the n sample observations consisting of the particular outcome reflected in the population proportion. This number of occurrence is denoted by X. Another sample statistic is the *sample proportion*, denoted by P, which is defined as

$$P = \frac{\textit{No. of elements with attribute } A}{\textit{Sample size}} = \frac{X}{n}$$

The observed value $p = x/n$ of sample proportion P will serve as an estimate of π. Obviously, the actual value we obtain for p will vary from sample to sample. So we ask, how good the estimate obtained will be. Are the values of P likely to be close to the true proportion π in the population. To what extent will they vary from one sample to another. Now for our theoretical model, we define a population in which a given proportion π have a specific attribute

A. We suppose that every unit in the population falls into one of the two categories A and $\bar{A}$. The notation is as follows

Number of units in A in	
Population	Sample
k	x

Proportion of units in A in	
Population	Sample
$\pi = \frac{k}{N}$	$p = \frac{x}{n}$

The estimate of proportion of successes π in the population is the sample proportion p and the estimate of the total number of successes k in population is thus Np or Nx/n.

11.14.3 Binomial Distribution as Sampling Distribution: ***Sampling Infinite Populations.*** If a simple random sample of size n is selected from an infinite population (or with replacement from a finite population) whose elements are characterised by some attribute to belong to one of the two mutually exclusive and exhaustive categories where one of these will be designated a 'success' and the other will be designated a 'failure', then the exact sampling distribution of the proportion of successes P is a binomial distribution.

11.14.4 Properties of Sampling Distribution of *P*. The properties of the sampling distribution of the sample proportion P are as follows:

Mean and Variance. The mean and variance of the binomial sampling distribution of P for a simple random sample of size n from an infinite Bernoulli population (or with replacement from a finite Bernoulli population) are given in the following theorem.

***Theorem* 11.11** *If the population is infinite or the sampling is done with replacement, the sample proportion P has its mean and variance as*

$$\mu_P = E(P) = \pi$$

$$\sigma_P^2 = Var(P) = \frac{\pi(1-\pi)}{n}$$

where π is the probability of success and $(1 - \pi)$ is the probability of failure. The standard deviation (often called the standard error or sampling variability) is

$$\sigma_P = \sqrt{Var(P)} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

However, if the value of π is unknown, it is replaced by sample proportion P, the estimate of the standard error of P then becomes

$$\hat{\sigma}_P = \sqrt{\frac{P(1-P)}{n}}$$

Shape of Distribution. The sampling distribution of P is skewed to the right if $\pi < 0.5$, skewed to the left if $\pi > 0.5$ and symmetrical if $\pi = 0.5$.

Normal approximation. As n tends to infinity, the distribution of P becomes approximately normal with mean π and variance $\pi(1-\pi)/n$. That is, the distribution of the random variable

$$Z = \frac{P - \mu_P}{\sigma_P} = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}}$$

approach the standard normal distribution as n approaches infinity.

***Example* 11.11** *A population consists of 5 members. The marital status of each member is given below*

Member	1	2	3	4	5
Marital status	*S*	*M*	*S*	*M*	*S*

where M and S stands for married and single respectively. Determine the proportion of married members in the population. Take all possible samples of two members with replacement from this population and find the proportion of married members in each sample. Form the sampling distribution of the sample proportion P and verify that

$$(i) \quad \mu_P = \pi \qquad (ii) \quad \sigma_P^2 = \frac{\pi(1-\pi)}{n}$$

Solution. Population: 1, 2, 3, 4, 5; Population size: $N = 5$; Sample size: $n = 2$

The members with even serial numbers 2 and 4 are married while those with odd serial numbers 1, 3 and 5 are single.

Number of married members in the population: $k = 2$

Proportion of married members in the population: $\pi = \frac{k}{N} = \frac{2}{5} = 0.4$

Number of possible samples $= N \times N = 5 \times 5 = 25$

All possible samples, the number of married members and the proportion of married members in each sample are given below.

Members in sample	Number of married members x	Proportion of married members $p = x/n$	Members in sample	Number of married members x	Proportion of married members $p = x/n$
1, 1	0	0	3, 3	0	0
1, 2	1	1/2	3, 4	1	1/2
1, 3	0	0	3, 5	0	0
1, 4	1	1/2	4, 1	1	1/2
1, 5	0	0	4, 2	2	1
2, 1	1	1/2	4, 3	1	1/2
2, 2	2	1	4, 4	2	1
2, 3	1	1/2	4, 5	1	1/2
2, 4	2	1	5, 1	0	0
2, 5	1	1/2	5, 2	1	1/2
3, 1	0	0	5, 3	0	0
3, 2	1	1/2	5, 4	1	1/2
		Continued	5, 5	0	0

The sampling distribution of sample proportion P, its mean and variance are

Value of P P	Number of occurrences f	Probability $f(p) = f/\Sigma f$	$p\,f(p)$	$p^2 f(p)$
0	9	9/25	0	0
1/2	12	12/25	6/25	3/25
1	4	4/25	4/25	4/25
Sum	$\Sigma f = 25$	1	10/25	7/25

$$\mu_P = E(P) = \Sigma p\, f(p) = \frac{10}{25} = 0.4$$

$$\sigma_P^2 = Var(P) = \Sigma p^2 f(p) - \mu_P^2 = \frac{7}{25} - (0.4)^2 = 0.12$$

We are to verify that (*i*) $\mu_P = \pi$ (*ii*) $\sigma_P^2 = \dfrac{\pi(1-\pi)}{n}$

$0.4 = 0.4$ $0.12 = \dfrac{0.4(1-0.4)}{2}$

$0.12 = 0.12$

***Example* 11.12** *It is known that* 3 % *of the persons living in Gujranwala city are known to have a certain disease. Find the mean and standard error of sampling distribution of proportion of diseased persons in a random sample of* 500 *persons.*

Solution. We have proportion in the population $\pi = 0.03$ and the sample size $n = 500$. Let P be the random variable 'the proportion of persons in the sample which are diseased'. Then, the mean and standard error of P are

$$\mu_P = \pi = 0.03$$

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.03(1-0.03)}{500}} = 0.00763$$

11.14.5 Hypergeometric Distribution as Sampling Distribution: ***Sampling Finite Populations.*** When a simple random sample of size n is selected without replacement from a finite population whose elements are characterised by some attribute to belong to one of the two mutually exclusive and exhaustive categories where one of these will be designated a 'success' and the other will be designated a 'failure', then the exact sampling distribution of the proportion of successes P is a hypergeometric distribution.

11.14.6 Properties of Sampling Distribution of *P*. The properties of the sampling distribution of the sample proportion P are as follows:

Mean and Variance. The mean and variance of the hypergeometric sampling distribution of P for simple random sampling without replacement from a finite Bernoulli population are given in the following theorem.

***Theorem* 11.12** *If the population is finite and the sampling is done without replacement, the sample proportion P has its mean and variance as*

$$\mu_P = E(P) = \pi$$

$$\sigma_P^2 = Var(P) = \frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right)$$

where π is the probability of success and $(1 - \pi)$ is the probability of failure. The standard deviation (often called the standard error or sampling variability) is

$$\sigma_P = \sqrt{Var(P)} = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

However, if the value of π is unknown, it is replaced by sample proportion P, the estimate of the standard error of P then becomes

$$\hat{\sigma}_P = \sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}$$

***Example* 11.13** *Draw all possible samples of size 2 at random without replacement from the population* 1, 2, 3, 4, 5. *Find the proportion of even numbers in the samples. Form the sampling distribution of the sample proportion P and verify that*

(*i*) $\mu_P = \pi$ (*ii*) $\sigma_P^2 = \frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right)$

Solution. Population: 1, 2, 3, 4, 5; Population size: $N = 5$; Sample size: $n = 2$

Number of even numbers in the population: $k = 2$

Proportion of even numbers in the population: $\pi = \frac{k}{N} = \frac{2}{5} = 0.4$

Number of possible samples $= N(N-1) = 5(5-1) = 20$

All possible samples, the number of even numbers and the proportion of even numbers in each sample are given below.

Sample values	Number of even numbers x	Proportion of even numbers $p = x/n$	Sample values	Number of even numbers x	Proportion of even numbers $p = x/n$
1, 2	1	1/2	3, 4	1	1/2
1, 3	0	0	3, 5	0	0
1, 4	1	1/2	4, 1	1	1/2
1, 5	0	0	4, 2	2	1
2, 1	1	1/2	4, 3	1	1/2
2, 3	1	1/2	4, 5	1	1/2
2, 4	2	1	5, 1	0	0
2, 5	1	1/2	5, 2	1	1/2
3, 1	0	0	5, 3	0	0
3, 2	1	1/2	5, 4	1	1/2

Continued

The sampling distribution of sample proportion P, its mean and variance are

Value of P	Number of occurrences	Probability		
p	f	$f(p) = f/\Sigma f$	$p\,f(p)$	$p^2 f(p)$
0	6	6/20	0	0
1/2	12	12/20	6/20	3/20
1	2	2/20	2/20	2/20
Sum	$\Sigma f = 20$	1	8/20	5/25

$$\mu_P = E(P) = \Sigma p\, f(p) = \frac{8}{20} = 0.4$$

$$\sigma_P^2 = Var(P) = \Sigma p^2 f(p) - \mu_P^2 = \frac{5}{20} - (0.4)^2 = 0.09$$

We are to verify that

(*i*) $\mu_P = \pi$ (*ii*) $\sigma_P^2 = \dfrac{\pi(1-\pi)}{n}\left(\dfrac{N-n}{N-1}\right)$

$0.4 = 0.4$ $0.12 = \dfrac{0.4(1-0.4)}{2}\left(\dfrac{5-2}{5-1}\right)$

$0.09 = 0.09$

11.15 SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS, $P_1 - P_2$

The *sampling distribution of the difference between two sample proportions* $P_1 - P_2$ is the probability distribution of all possible differences between proportions P_1 and P_2 obtained from all possible independent simple random samples of n_1 and n_2 observations that can be drawn from two Bernoulli populations with population proportions of π_1 and π_2, respectively. Often we wish to compare the proportions of successes in two Bernoulli populations. We must use the sample proportions of successes as our basis of comparison. Obviously, the number of successes in both samples cannot be used alone as a means of evaluation. Specifically we require a probability model of the difference between two sample proportions.

Suppose that two independent random samples of sizes n_1 and n_2 are drawn from Bernoulli populations with population proportions of π_1 and π_2, respectively. Let P_1 be the proportion of successes in sample of size n_1 from the population with population proportion π_1, then P_1 is a random variable that has its own probability distribution with mean π_1 and variance $\pi_1(1-\pi_1)/n_1$. Let P_2 be the proportion of successes in sample of size n_2 from the population with population proportion π_2, then P_2 is a random variable that has its own probability distribution with mean π_2 and variance $\pi_2(1-\pi_2)/n_2$. Then the difference $P_1 - P_2$ can be obtained from all possible pairs of P_1 and P_2. Consequently, the difference $P_1 - P_2$ between the

two sample proportions is a random variable that has its own probability distribution which is called the sampling distribution of the difference between two sample proportions.

11.15.1 Properties of the Sampling Distribution of the Difference between Two Sample Proportions. The properties of the sampling distribution of the difference $P_1 - P_2$ between two sample proportions are given by the following theorems.

***Theorem* 11.13** *The mean of the sampling distribution of* $P_1 - P_2$, *denoted by* $\mu_{P_1 - P_2}$, *is equal to the difference between the population proportions, i. e.,*

$$\mu_{P_1 - P_2} = E(P_1 - P_2) = \pi_1 - \pi_2$$

This theorem holds regardless of the sample sizes n_1 and n_2 or whether the sampling is done with or without replacement.

***Theorem* 11.14** *If the populations are infinite or the sampling is done with replacement, the difference between sample proportions* $P_1 - P_2$ *has its variance as*

$$\sigma^2_{P_1 - P_2} = Var(P_1 - P_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

The standard error of $P_1 - P_2$ becomes

$$\sigma_{P_1 - P_2} = \sqrt{Var(P_1 - P_2)} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

However, if the values of π_1 and π_2 are unknown, these are replaced by sample proportions P_1 and P_2, the estimate of the standard error of $P_1 - P_2$ then becomes

$$\hat{\sigma}_{P_1 - P_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

***Theorem* 11.15** *If the populations are finite and the sampling is done without replacement, the difference between sample proportions* $P_1 - P_2$ *has its variance as*

$$\sigma^2_{P_1 - P_2} = Var(P_1 - P_2)$$

$$= \frac{\pi_1(1-\pi_1)}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\pi_2(1-\pi_2)}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)$$

The standard error of $P_1 - P_2$ is

$$\sigma_{P_1 - P_2} = \sqrt{Var(P_1 - P_2)}$$

$$= \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}\left(\frac{N_2 - n_2}{N_2 - 1}\right) + \frac{\pi_2(1-\pi_2)}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

***Example* 11.14** *Let P_1 represent the proportion of odd numbers in a sample of size $n_1 = 2$ selected at random with replacement from a finite population consisting of values 4 and 5. Similarly, let P_2 represent the proportion of odd numbers in a sample of size $n_2 = 2$ selected at random with replacement from another finite population consisting of values 2, 3 and 6. From a sampling distribution of the random variable $(P_1 - P_2)$. Verify that*

(*i*) $\mu_{P_1 - P_2} = \pi_1 - \pi_2$ (*ii*) $\sigma^2_{P_1 - P_2} = \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}$

Solution. We have

***Population* I:** 4, 5; $N_1 = 2$; $n_1 = 2$

Number of odd numbers: $k_1 = 2$

Proportion of odd numbers: $\pi_1 = \dfrac{k_1}{N_1} = \dfrac{1}{2}$

Number of possible samples: $N_1 \times N_1 = 2 \times 2 = 4$

Possible samples:

(4, 4) (4, 5) (5, 4) (5, 5)

Sample proportion of odd numbers: p_1

0 1/2 1/2 1

***Population* II:** 2, 3, 6; $N_2 = 3$; $n_2 = 2$

Number of odd numbers: $k_2 = 1$

Proportion of odd numbers: $\pi_2 = \dfrac{k_2}{N_2} = \dfrac{1}{3}$

Number of possible samples: $N_2 \times N_2 = 3 \times 3 = 9$

Possible samples:

(2, 2) (2, 3) (2, 6) (3, 2) (3, 3) (3, 6) (6, 2) (6, 3) (6, 6)

Sample proportion of odd numbers: p_2

0 1/2 0 1/2 1 1/2 0 1/2 0

All possible differences between sample proportions $(P_1 - P_2)$ are

					p_2				
p_1	0	0	0	0	1/2	1/2	1/2	1/2	1
0	0	0	0	0	−1/2	−1/2	−1/2	−1/2	−1
1/2	1/2	1/2	1/2	1/2	0	0	0	0	−1/2
1/2	1/2	1/2	1/2	1/2	0	0	0	0	−1/2
1	1	1	1	1	1/2	1/2	1/2	1/2	0

The sampling distribution of $P_1 - P_2$, its mean and variance are

Value of $P_1 - P_2$	Number of occurrences	Probability		
$p_1 - p_2$	f	$f(p_1 - p_2) = f / \Sigma f$	$(p_1 - p_2) f(p_1 - p_2)$	$(p_1 - p_2)^2 f(p_1 - p_2)$
-1	1	$1/36$	$-1/36$	$1/36$
$-1/2$	6	$6/36$	$-3/36$	$3/72$
0	13	$13/36$	0	0
$1/2$	12	$12/36$	$6/36$	$3/36$
1	4	$4/36$	$4/36$	$4/36$
Sum	$\Sigma f = 36$	1	$6/36$	$19/72$

$$\mu_{P_1 - P_2} = E(P_1 - P_2) = \Sigma (p_1 - p_2)(p_1 - p_2) = \frac{6}{36} = \frac{1}{6}$$

$$\sigma^2_{P_1 - P_2} = Var(P_1 - P_2) = \Sigma (p_1 - p_2)^2 f(p_1 - p_2) - \mu^2_{P_1 - P_2}$$

$$= \frac{19}{72} - \left(\frac{1}{6}\right)^2 = \frac{17}{72}$$

We are to verify that

(*i*) $\mu_{P_1 - P_2} = \pi_1 - \pi_2$ (*ii*) $\sigma^2_{P_1 - P_2} = \dfrac{\pi_1 (1 - \pi_1)}{n_1} + \dfrac{\pi_2 (1 - \pi_2)}{n_2}$

$$\frac{1}{6} = \frac{1}{2} - \frac{1}{3} \qquad \frac{17}{72} = \frac{\frac{1}{2}\left(1 - \frac{1}{2}\right)}{2} + \frac{\frac{1}{3}\left(1 - \frac{1}{3}\right)}{2}$$

$$\frac{1}{6} = \frac{1}{6} \qquad \frac{17}{72} = \frac{17}{72}$$

***Example* 11.15** *The actual proportion of men who like a certain TV programme is* 0.30 *and the corresponding proportion for women is* 0.25. *A questionnaire about this program is given to* 500 *men and* 500 *women, and the individual responses are looked upon as the values of independent random variables having Bernoulli distributions with parameters* $\pi_1 = 0.30$ *and* $\pi_2 = 0.25$, *respectively. Find the mean and standard error of* $P_1 - P_2$, *the difference between the sample proportions of successes.*

Solution. We have $\pi_1 = 0.30$, $\pi_2 = 0.25$; $n_1 = 500$, $n_2 = 500$.

The mean and standard error of $P_1 - P_2$ are

$$\mu_{P_1 - P_2} = \pi_1 - \pi_2 = 0.30 - 0.25 = 0.05$$

$$\sigma_{P_1 - P_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

$$= \sqrt{\frac{0.30(1-0.30)}{500} + \frac{0.25(1-0.25)}{500}} = 0.028$$

11.16 OTHER SAMPLING DISTRIBUTIONS

We have considered the sampling distributions of sample mean, difference between sample means, sample proportions and difference between sample proportions. Other statistics such as sample median, sample variance and sample standard deviation have their own sampling distributions. There is different sampling distribution for each different statistic even though the statistics may be computed from the same sample. For a given statistic, the sampling distribution will vary for samples of different sizes. Thus, in a sampling distribution it is necessary to specify the population, the statistic and the size of the sample. A change in any of these specifications will result a different sampling distribution.

11.17 SAMPLING DISTRIBUTION OF THE SAMPLE VARIANCE, S^2

The *sampling distribution of sample variance* S^2 is the probability distribution of the variances obtained from all possible simple random samples of n observations that can be drawn from a population with variance σ^2.

The sampling distribution of sample variance has the property

$$\mu_{S^2} = E(S^2) = \frac{n-1}{n}\sigma^2$$

***Example* 11.16** *A population consists of five numbers* 2, 4, 6, 8, *and* 10. *Consider all possible samples of size* 2 *which can be drawn with replacement from this population. Form the sampling distribution of sample variance and verify that*

$$\mu_{S^2} = \frac{n-1}{n}\sigma^2$$

Solution. Population: 2, 4, 6, 8, 10; Population size: $N = 5$; Sample size : $n = 2$

Number of possible samples $= N \times N = 5 \times 5 = 25$

All possible samples that can be drawn with replacement from our population are

(2, 2)	(2, 4)	(2, 6)	(2, 8)	(2, 10)
(4, 2)	(4, 4)	(4, 6)	(4, 8)	(4, 10)
(6, 2)	(6, 4)	(6, 6)	(6, 8)	(6, 10)
(8, 2)	(8, 4)	(8, 6)	(8, 8)	(8, 10)
(10, 2)	(10, 4)	(10, 6)	(10, 8)	(10, 10)

(*i*) All possible sample variances: $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{(x_1 - x_2)^2}{4}$ when $n = 2$

0	1	4	9	16
1	0	1	4	9
4	1	0	1	4
9	4	1	0	1
16	9	4	1	0

The sampling distribution of sample variance S^2 and its mean are

Value of S^2 s^2	Number of occurrences: f	Probability $p(s^2) = f/\Sigma f$	$s^2\, p(s^2)$
0	5	5/25	0
1	8	8/25	8/25
4	6	6/25	24/25
9	4	4/25	36/25
16	2	2/25	32/25
	$\Sigma f = 25$	1	$\Sigma s^2 p(s^2) = 100/25$

$$\mu_{S^2} = E(S^2) = \Sigma s^2\, p(s^2) = \frac{100}{25} = 4$$

The mean and variance of the population are

x_j	2	4	6	8	10	$\Sigma x_j = 30$
x_j^2	4	16	36	64	100	$\Sigma x_j^2 = 220$

$$\mu = \frac{\Sigma x_j}{N} = \frac{30}{5} = 6$$

$$\sigma^2 = \frac{\Sigma x_j^2}{N} - \mu^2 = \frac{220}{5} - (6)^2 = 8$$

We are to verify that $\mu_{S^2} = \dfrac{n-1}{n}\sigma^2$

$$4 = \frac{2-1}{2}(8)$$

$$4 = 4$$

Exercise 11.4

1. ***(a)*** A fair coin is tossed 50 times and the number of heads recorded are 27. The proportion of heads was, therefore, estimated to be 0.54. Answer the following.

(*i*) Which figure is a parameter?

(*ii*) Which figure is a statistic?

{ (*i*) The probability of head in a single trial $\pi = 0.5$;

(*ii*) Sample size $n = 50$, number of heads in the sample $x = 30$ and the proportion of heads in the sample $p = x/n = 0.54$. }

(b) What is meant by the sampling distribution of sample proportion? Describe the properties of the sampling distribution of sample proportion.

2. (*a*) A finite population consists of the numbers 2, 3, 4, 5, 6 and 8. Find the proportion P of even numbers in all possible random samples of size $n = 2$ that can be drawn with replacement from this population. Assuming the 36 possible samples equally likely, make the sampling distribution of sample proportions and find the mean and variance of this distribution. Verify that

(*i*) $E(P) = \pi$ (*ii*) $Var(P) = \dfrac{\pi(1-\pi)}{n}$

where P and π are sample and population proportions respectively.

$\{\pi = 2/3,\ \mu_P = 2/3,\ \sigma_P^2 = 1/9\}$

(*b*) A population consists of $N = 4$ numbers 1, 3, 4 and 5. Find the proportion P of odd numbers in all possible samples of size $n = 3$ that can be drawn without replacement from this population. Assuming the 24 possible samples equally likely, construct the sampling distribution of sample proportions and find the mean and variance of this distribution. Verify that

(*i*) $\mu_P = \pi$, (*ii*) $\sigma_P^2 = \dfrac{\pi(1-\pi)}{n}\left(\dfrac{N-n}{N-1}\right)$

where P and π are sample and population proportions respectively.

$\{\pi = 3/4,\ \mu_P = 3/4,\ \sigma_P^2 = 1/48\}$

3. (*a*) Suppose that 60 % of a city population favours public finding for a proposed recreational facility. If 150 persons are to be randomly selected and interviewed, what is the mean and standard error of the sample proportion favouring this issue.

$\{\mu_P = 0.60,\ \sigma_P = 0.04\}$

(*b*) A small, professional society has $N = 4500$ members. The president has mailed $n = 400$ questionnaires to a random sample of members asking whether they wish to affiliate with a large group. Assuming that the proportion of the entire membership favouring consolidation is $\pi = 0.7$, find the mean and standard error of the sample proportion P.

$\{\mu_P = 0.7,\ \sigma_P = 0.022\}$

4. (*a*) What is meant by the sampling distribution of the difference between two sample proportions? Describe the properties of the sampling distribution of difference between two sample proportions. Explain its usefulness in statistical inference.

(*b*) Let P_1 represent the proportion of odd numbers in a random sample of size $n_1 = 3$ with replacement from a finite population consisting of values 4 and 5. Similarly, let P_2 represent the proportion of odd numbers in a random sample of size $n_2 = 2$ with replacement from another finite population consisting of values 2, 3 and 6. Assuming that the 72 possible differences $P_1 - P_2$ are equally likely to occur, construct the sampling distribution of $P_1 - P_2$. Verify that

(*i*) $\mu_{P_1 - P_2} = \pi_1 - \pi_2$ (*ii*) $\sigma^2_{P_1 - P_2} = \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}$

$\{\pi_1 = 1/2,\ \pi_2 = 1/3,\ \mu_{P_1 - P_2} = 1/6,\ \sigma^2_{P_1 - P_2} = 7/36\}$

5. (*a*) Let P_1 represent the proportion of even numbers in a random sample of size $n_1 = 2$ without replacement from a finite population consisting of values 4, 6 and 9. Similarly, let P_2 represent the proportion of even numbers in a random sample of size $n_2 = 2$ without replacement from another finite population consisting of values 2, 3 and 5. Assuming that the 36 possible differences $P_1 - P_2$ are equally likely to occur, construct the sampling distribution of $P_1 - P_2$. Verify that

(*i*) $\mu_{P_1 - P_2} = \pi_1 - \pi_2$

(*ii*) $\sigma^2_{P_1 - P_2} = \dfrac{\pi_1(1-\pi_1)}{n_1}\left(\dfrac{N_1 - n_1}{N_1 - 1}\right) + \dfrac{\pi_2(1-\pi_2)}{n_2}\left(\dfrac{N_2 - n_2}{N_2 - 1}\right)$

$\{\pi_1 = 2/3,\ \pi_2 = 1/3,\ \mu_{P_1 - P_2} = 1/3,\ \sigma^2_{P_1 - P_2} = 1/9\}$

(*b*) The percentage of families with a monthly income of Rs. 1,000 or more in city *A* and city *B* is 25 % and 20 % respectively. If a random sample of 100 families is selected from each of these two cities and the proportions of families earning Rs. 1,000 or more in the two samples are compared, what is the mean and standard error of $P_1 - P_2$, the difference between the sample proportions?

$\{\mu_{P_1 - P_2} = 0.05,\ \sigma_{P_1 - P_2} = 0.059\}$

6. (*a*) A finite population consists of five values 2, 4, 6, 8 and 10. Take all possible samples of size 2 which can be drawn with replacement from this population. Assuming the 25 possible samples equally likely, construct the sampling distributions of sample means and sample variances and find the mean of these distributions. Calculate the mean and variance of the population and verify that

(*i*) $\mu_{\bar{x}} = \mu$ (*ii*) $\mu_{s^2} = \dfrac{n-1}{n}\sigma^2$

where $\bar{X} = \dfrac{\Sigma X_i}{n}$ and $S^2 = \dfrac{\Sigma(X_i - \bar{X})^2}{n}$

$\{\mu = 6,\ \sigma^2 = 8,\ \mu_{\bar{x}} = 6,\ \mu_{s^2} = 4\}$

(*b*) A finite population consists of five values 1, 3, 5, 7 and 9. Take all possible samples of size 2 which can be drawn with replacement from this population. Assuming the 25 possible samples equally likely, construct the sampling distributions of sample means and sample variances and find the mean of these distributions. Calculate the mean and variance of the population. Discuss the results.

$\left(\mu = 5,\ \sigma^2 = 8,\ \mu_{\bar{x}} = 5,\ \mu_{s^2} = 4,\ \mu_{\bar{x}} = \mu,\ \mu_{s^2} = \dfrac{n-1}{n}\sigma^2\right)$

7. (*a*) A finite population consists of 5 values 1, 3, 5, 7 and 9. Take all possible samples of size 2 which can be drawn without replacement from this population. Assuming the 20 possible samples equally likely, construct the sampling distributions of sample means and sample variances and find the mean of these distributions. Calculate the mean

and variance of the population and verify that

(*i*) $\mu_{\bar{x}} = \mu$ (*ii*) $\mu_{s^2} = \frac{N}{N-1}\frac{n-1}{n}\sigma^2$

where $\bar{X} = \frac{\Sigma X_i}{n}$ and $S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n}$

$\{\mu = 5, \sigma^2 = 8, \mu_{\bar{x}} = 5, \mu_{s^2} = 5\}$

(*b*) Take all possible samples of 2 distinct values from the population 2, 4, 6, 8 and 10. Assuming the 20 possible samples equally likely, construct the sampling distributions of sample means and sample variances and find the mean of these distributions. Calculate the mean and variance of the population. Discuss the results.

$\left(\mu = 6, \sigma^2 = 8, \mu_{\bar{x}} = 6, \mu_{s^2} = 5, \mu_{\bar{x}} = \mu, \mu_{s^2} = \frac{N}{N-1}\frac{n-1}{n}\sigma^2\right)$

◆

Exercise 11.5
Objective Questions

1. Fill in the blanks.

(*i*) A ——— is the totality of the observations made on all the objects possessing some common specific characteristics. (*population*)

(*ii*) A ——— is a part of the population which is selected with the expectation that it will represent the characteristics of the population. (*sample*)

(*iii*) ——— is a procedure of selecting a representative sample from a given population. (*Sampling*)

(*iv*) The descriptive measures of a population are called ———. (*parameters*)

(*v*) A descriptive measure on the sample observations is called ———. (*statistic*)

(*vi*) A population is called ——— if it includes a limited number of sampling units. (*finite*)

(*vii*) A population is called ——— if it includes an unlimited number of sampling units. (*infinite*)

(*viii*) Sampling ——— is a complete list of the sampling units. (*frame*)

(*ix*) A ——— sampling is a procedure in which we cannot assign to an element of the population the probability of its being included in the sample. (*non-probability*)

(*x*) A ——— sampling is a process in which the sample is selected in such a way that every element of a population has a known nonzero probability of being included in the sample. (*probability*)

(*xi*) Another name of a probability sampling is ——— sampling. (*random*)

(*xii*) Random sampling provides reliable ————. (*estimates*)

(*xiii*) The sampling is said to be ———— replacement when the unit selected at random is returned to the population before the next unit is selected. (*with*)

(*xiv*) The sampling is said to be ———— replacement when the unit selected at random is returned to the population before the next unit is selected. (*without*)

(*xv*) A sample is usually selected by ———— replacement. (*without*)

(*xvi*) In sampling ———— replacement, a sampling unit can be selected more than once. (*with*)

2. Fill in the blanks.

(*i*) In sampling ———— replacement, a sampling unit cannot be selected more than once. (*without*)

(*ii*) In sampling with replacement, a finite population becomes ————. (*infinite*)

(*iii*) ———— random sampling is a procedure of selecting a sample from the population in such a way that every unit available for sampling has an equal probability of being selected. (*simple*)

(*iv*) The sampling error decreases by increasing the sample ————. (*size*)

(*v*) The ———— errors may be present both in sample survey and census. (*non-sampling*)

(*vi*) The bias increases by increasing the sample ————. (*size*)

(*vii*) A sample which is free from bias is called an ———— sample. (*unbiased*)

(*viii*) ———— errors may arise due to faulty sampling frames, non-responses and processing of data. (*Non-sampling*)

(*ix*) ———— errors can be controlled by the proper training of the investigators and following up the non-responses (*Non-sampling*)

(*x*) The probability distribution of a sample statistic is called ———— distribution of that statistic. (*sampling*)

(*xi*) The standard deviation of sampling distribution of a sample statistic is called the ———— of that statistic. (*standard error*)

(*xii*) The standard error can be reduced by increasing the ————. (*sample size*)

(*xiii*) The number of all possible samples of size n taken with replacement from a population of size N is ————. $(N)^n$

(*xiv*) The number of all possible samples of size n taken without replacement from a population of size N is ————. $({}^{N}P_{n})$

3. Mark off the following statements as true or false.

(*i*) A descriptive measure on the sample observations is called parameter and a descriptive measure of a population is called statistic. (*false*)

(*ii*) A sample statistic is a random variable whereas the parameter being estimated is constant. (*true*)

(*iii*) A sample survey provides the results which are more accurate than those obtained from a census. (*false*)

(*iv*) A sample design is a procedure for obtaining a sample from a given population prior to collecting any data. (*true*)

(*v*) More detailed information can be obtained in a sample survey as compared to a census. (*true*)

(*vi*) Sampling may be the only means available for obtaining the desired information if the population is infinite. (*true*)

(*vii*) If the data are obtained by tests that are destructive, then complete enumeration becomes essential. (*false*)

(*viii*) Every random sample is a simple random sample. (*false*)

(*ix*) In sampling with replacement, the sample size may be greater than the population size. (*true*)

(*x*) In sampling without replacement, the sample size can be greater than population size. (*false*)

(*xi*) The number of units available for the next drawing does not change in a random sampling with replacement. (*true*)

(*xii*) In sampling without replacement, the number of units remaining after each drawing will be reduced by one. (*true*)

4. Mark off the following statements as true or false.

(*i*) The number of all possible samples of size n taken without replacement from a population of size N is ${}^{N}C_{n}$. (*false*)

(*ii*) In sampling without replacement, a sampling unit can be selected more than once. (*false*)

(*iii*) In sampling with replacement, the sample size may be greater than the population size. (*true*)

(*iv*) In sampling with replacement, a finite population becomes infinite. (*true*)

(*v*) Non-sampling errors may be present both in sample survey and census. (*true*)

(*vi*) The sampling error increases by increasing the sample size. (*false*)

(*vii*) Sampling and non-sampling errors are both controllable. (*true*)

(*viii*) The standard deviation of a sampling distribution of a statistic is called the standard error of that statistic. (*true*)

(*ix*) Standard error is the difference of a statistic from the parameter being estimated. (*false*)

(*x*) We can decrease both sampling error and standard error by increasing the sample size. (*true*)

(*xi*) The reliability of an estimate can be determined by its standard error. (*true*)

12 ESTIMATION

12.1 STATISTICAL INFERENCE

Statistical inference is a field concerned with drawing conclusions about distributions by using observed values of random variables which are governed by these distributions.

Statistical inferences are the conclusions made about the unknown value of the parameter of a population using a limited information contained in an observed sample taken from it at random. The two most important types of statistical inferences are

(*i*) Estimation of parameters

(*ii*) Testing of hypotheses

12.2 STATISTICAL ESTIMATION

The *statistical estimation* is a procedure of making judgment about the unknown value of a population parameter by using the sample observations.

Population parameters are estimated from sample data because it is impracticable to examine the entire population in order to make such an exact determination. Statistical estimation procedures provide estimates of population parameters with a desired degree of confidence. This degree of confidence can be controlled, in part, by the size of the sample (the larger the sample, the greater the accuracy of the estimate) and by the type of the estimate made. The statistical estimation of population parameters is further divided into two types

(*i*) Point estimation

(*ii*) Interval estimation

12.3 POINT ESTIMATION OF A PARAMETER

The object of *point estimation* is to obtain a single number from the sample that is intended for estimating the unknown true value of a population parameter.

12.3.1 Point Estimator. A *point estimator* is a sample statistic that is used to estimate the unknown true value of a population parameter.

An estimator is always a statistic which is both a function and random variable with a probability distribution. An estimator is denoted by a capital letter (*e. g.*, T, U, $\cdots$).

Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from a population with probability mass function or probability density function $f(x\,;\,\theta)$, then the estimator T intended to estimate θ is a function given by

$$T = g(X_1, X_2, \cdots, X_n)$$

12.3.2 Point Estimate. A *point estimate* is a specific value of an estimator computed from the sample data after the sample has been observed. When a random sample becomes available from

the population and the estimator T is computed from the sample data, the numerical value obtained is an estimate of population parameter θ from the particular sample. An estimate is denoted by a small letter (*e. g.*, t, u, $\cdots$).

Suppose that $x_1, x_2, \cdots, x_n$ is an observed random sample from a population with probability mass function or probability density function $f(x; \theta)$, then the particular value of an estimator T intended to estimate θ is a given by

$$t = g(x_1, x_2, \cdots, x_n)$$

***Example* 12.1** *A random sample selected from a normal population with mean μ and variance σ^2 gave the values* 25, 31, 23, 33, 28, 36, 22, 26. *Give the point estimators for μ and σ^2 and find their point estimates.*

Solution. We have

x_i	25	31	23	33	28	36	22	26	$\Sigma x_i = 224$
$x_i - \bar{x}$	-3	3	-5	5	0	8	-6	-2	
$(x_i - \bar{x})^2$	9	9	25	25	0	64	36	4	$\Sigma(x_i - \bar{x})^2 = 172$

The point estimator of population mean μ is $\quad \bar{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$

The point estimate of population mean μ is $\quad \bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{224}{8} = 28$

The point estimators of population variance σ^2 are

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2, \qquad \hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The point estimates of population variance σ^2 are

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{172}{8} = 21.5, \qquad \hat{s}^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{172}{8-1} = 24.57$$

12.4 UNBIASEDNESS

The distribution of an estimator should be centred in some sense at the value of the parameter to be estimated. Because expected value is a measure of the centre of a distribution, a reasonable requirement for an estimator T may be $E(T) = \theta$. This property is called as unbiasedness of the estimator T. It refers to the desirability of the sampling distribution of an estimator being centred at the parameter to be estimated.

12.4.1 Unbiased Estimator. An estimator is *unbiased* if the mean of its sampling distribution is equal to the population parameter to be estimated.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution $f(x; \theta)$. An estimator $T = g(X_1, X_2, \cdots, X_n)$ is said to be *unbiased* for parameter θ if

$$E(T) = \theta$$

12.4.2 Biased Estimator. An estimator T of a population parameter θ is said to be biased if:

$$E(T) \neq \theta$$

12.4.3 Bias. If an estimator T of a population parameter θ is biased, the amount of its bias is

$$Bias = E(T) - \theta$$

If T is an unbiased estimator, it will tend to give estimates nearer to θ and if T is a biased estimator, it will tend to give estimates far from θ.

***Example* 12.2** *A population consists of five numbers* 2, 4, 6, 8, *and* 10. *Consider all possible samples of size* 2 *which can be drawn with replacement from this population. By forming the sampling distributions, show that*

(*i*) *The sample variance* $S^2 = \sum(X_i - \bar{X})^2/n$ *is a biased estimator of the population variance* σ^2.

(*ii*) *The sample variance* $\hat{S}^2 = \sum(X_i - \bar{X})^2/(n-1)$ *is an unbiased estimator of the population variance* σ^2.

Solution. Population: 2, 4, 6, 8, 10 Population size: $N = 5$ Sample size : $n = 2$

The mean and variance of the population are

x_j	2	4	6	8	10	$\sum x_j = 30$
x_j^2	4	16	36	64	100	$\sum x_j^2 = 220$

$$\mu = \frac{\sum x_j}{N} = \frac{30}{5} = 6$$

$$\sigma^2 = \frac{\sum x_j^2}{N} - \mu^2 = \frac{220}{5} - (6)^2 = 8$$

Number of possible samples $= N \times N = 5 \times 5 = 25$

All possible samples:

(2, 2)	(2, 4)	(2, 6)	(2, 8)	(2, 10)
(4, 2)	(4, 4)	(4, 6)	(4, 8)	(4, 10)
(6, 2)	(6, 4)	(6, 6)	(6, 8)	(6, 10)
(8, 2)	(8, 4)	(8, 6)	(8, 8)	(8, 10)
(10, 2)	(10, 4)	(10, 6)	(10, 8)	(10, 10)

(*i*) All possible sample variances: $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n} = \dfrac{(x_1 - x_2)^2}{4}$ when $n = 2$

0	1	4	9	16
1	0	1	4	9
4	1	0	1	4
9	4	1	0	1
16	9	4	1	0

The sampling distribution S^2 and its mean are

Value of S^2 s^2	Number of occurrences f	Probability $p(s^2) = f/\Sigma f$	$s^2\ p(s^2)$
0	5	5/ 25	0
1	8	8/ 25	8/ 25
4	6	6/ 25	24/ 25
9	4	4/ 25	36/ 25
16	2	2/ 25	32/ 25
	$\Sigma f = 25$	1	$\Sigma s^2 p(s^2) = 100/25$

$$E(S^2) = \Sigma s^2\, p(s^2) = \frac{100}{25} = 4$$

Since $4 = E(S^2) \neq \sigma^2 = 8$, therefore S^2 is a biased estimator of σ^2.

(ii) All possible sample variances: $\hat{s}^2 = \dfrac{\Sigma(x_i - \bar{x})^2}{n-1} = \dfrac{(x_1 - x_2)^2}{2}$ when $n = 2$

0	2	8	18	32
2	0	2	8	18
8	2	0	2	8
18	8	2	0	2
32	18	8	2	0

The sampling distribution of $\hat{S}^2$ and its mean are

Value of $\hat{S}^2$ $\hat{s}^2$	Number of occurrences f	Probability $p(\hat{s}^2) = f/\Sigma f$	$\hat{s}^2\ p(\hat{s}^2)$
0	5	5/ 25	0
2	8	8/ 25	16/ 25
8	6	6/ 25	48/ 25
18	4	4/ 25	72/ 25
32	2	2/ 25	64/ 25
	$\Sigma f = 25$	1	$\Sigma \hat{s}^2 p(\hat{s}^2) = 200/25$

$$E(\hat{S}^2) = \Sigma \hat{s}^2\, p(\hat{s}^2) = \frac{200}{25} = 8$$

Since $8 = E(\hat{S}^2) = \sigma^2 = 8$, therefore $\hat{S}^2$ is an unbiased estimator of σ^2.

12.5 BEST ESTIMATOR

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from the distribution $f(x; \theta)$. Among the class U of all unbiased estimators $T = g(X_1, X_2, \cdots, X_n)$ for a given parameter θ, the estimator T^* is said to be a *best* or *minimum variance* estimator if among the class U of all unbiased estimators, none has a smaller variance than T^*.

12.5.1 Best Estimators of the Population Mean and Variance. Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population with unknown mean μ and unknown variance σ^2, then the best estimators of μ and σ^2 are

$$\bar{X} = \frac{\Sigma X_i}{n} \qquad \text{and} \qquad \hat{S}^2 = \frac{\Sigma (X_i - \bar{X})^2}{n - 1}$$

respectively.

***Example* 12.3** *Obtain the best unbiased estimates of the population mean μ and variance σ^2 from which the following sample is drawn:*

$$n = 8, \qquad \Sigma x_i = 120, \qquad \Sigma (x_i - \bar{x})^2 = 302$$

Solution. The best estimate of the population mean μ is the sample mean

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{120}{8} = 15$$

The best estimate of the population variance σ^2 is the sample variance

$$\hat{s}^2 = \frac{\Sigma (x_i - \bar{x})^2}{n - 1} = \frac{302}{8 - 1} = 43.14$$

12.5.2 Best estimator of the population proportion. From a population which has unknown proportion of successes π, we take a random sample of size n with X as the number of successes in the sample, then the best estimator of π is

$$P = \frac{X}{n}$$

***Example* 12.4** *A random sample of 50 children from a large school is chosen and the number who are left handed is noted. It is found that 6 are left handed. Obtain an unbiased estimate of the proportion of children in the school who are left handed.*

Solution. We have $n = 50 \qquad x = 6$

Sample proportion: $p = \frac{x}{n} = \frac{6}{50} = 0.12$

12.6 POOLED ESTIMATORS FROM TWO SAMPLES

Estimates of the population mean, variance, proportion, *etc.*, may be obtained by pooling observations from two random samples.

12.6.1 Pooled Estimator of Population Mean. Let $X_{11}, X_{21}, \cdots, X_{n_1 1}$ and $X_{12}, X_{22}, \cdots, X_{n_2 2}$ be two random samples of sizes n_1 and n_2 from a population with unknown mean μ, then the pooled estimator $\overline{X}_p$ of μ is

$$\overline{X}_p = \frac{\sum_{i=1}^{n_1} X_{i1} + \sum_{i=1}^{n_2} X_{i2}}{n_1 + n_2} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the unbiased estimators of μ, based on the first and the second sample, respectively.

12.6.2 Pooled Estimator of Population Variance. Let $X_{11}, X_{21}, \cdots, X_{n_1 1}$ and $X_{12}, X_{22}, \cdots, X_{n_2 2}$ be two random samples of sizes n_1 and n_2 from a population with unknown variance σ^2, then the pooled estimator S_p^2 of σ^2 is

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{i1} - \overline{X}_1)^2 + \sum_{i=1}^{n_2}(X_{i2} - \overline{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\,\hat{S}_1^2 + (n_2 - 1)\,\hat{S}_2^2}{n_1 + n_2 - 2}$$

where $\hat{S}_1^2$ and $\hat{S}_2^2$ are the unbiased estimators of σ^2, based on the first and the second sample, respectively.

***Example* 12.5** *Two samples of sizes* 40 *and* 50, *respectively, are taken from a population with unknown mean* μ *and unknown variance* σ^2.

Sample I: $n_1 = 40, \quad \Sigma f x_1 = 807, \quad \Sigma f x_1^2 = 16329$

Sample II: $n_2 = 50, \quad \Sigma f x_2 = 977, \quad \Sigma f x_2^2 = 19177$

Using the data from the two samples, obtain the best estimates of μ *and* σ^2.

Solution. The best estimates of μ and σ^2 are

$$\bar{x}_p = \frac{\Sigma f x_1 + \Sigma f x_2}{n_1 + n_2} = \frac{807 + 977}{40 + 50} = 19.82$$

$$\Sigma f(x_1 - \bar{x}_1)^2 = \Sigma f x_1^2 - \frac{(\Sigma f x_1)^2}{n_1} = 16329 - \frac{(807)^2}{40} = 47.775$$

$$\Sigma f(x_2 - \bar{x}_2)^2 = \Sigma f x_2^2 - \frac{(\Sigma f x_2)^2}{n_2} = 19177 - \frac{(977)^2}{50} = 86.42$$

$$s_p^2 = \frac{\Sigma f(x_1 - \bar{x}_1)^2 + \Sigma f(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{47.775 + 86.42}{40 + 50 - 2} = 1.525$$

12.6.3 Pooled Estimator of Population Proportion. From a population which has unknown proportion of successes π, we take two random samples of sizes n_1 and n_2 with X_1 and X_2 as the number of successes in the respective sample, then the pooled estimator $\hat{\pi}$ of π is

$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

where P_1 and P_2 are the unbiased estimators of π, based on the first and second sample, respectively.

***Example* 12.6** *A random sample of* 600 *people from a certain district were questioned and the results indicated that* 30% *used a particular product. In a second random sample of* 300 *people,* 96 *used the product. Using the data from the two samples, find the best estimate of the proportion of people in the district who used the product.*

Solution. The best estimate of the population proportion is

$$n_1 = 600 \qquad p_1 = 0.30$$

$$n_2 = 300 \qquad x_2 = 96 \qquad \Rightarrow \qquad p_2 = \frac{x_2}{n_2} = \frac{96}{300} = 0.32$$

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{600(0.30) + 300(0.32)}{600 + 300} = 0.307$$

Exercise 12.1

1. (*a*) Explain what is meant by statistical inference?

(*b*) What is meant by estimation? Differentiate between estimator and estimate?

2. (*a*) Specify the estimator and the estimate in each of the following:

(*i*) A sample of 35 students gave an average height of 62 inches.

(*ii*) A sample of 50 households having television sets showed that 85 percent of them liked a particular programme.

(*iii*) A sample of 25 bolts produced by a company showed that 20 of them were according to specifications.

(*iv*) A sample of 30 houses showed an average consumption of electricity as 65 units.

{ (*i*) Sample mean height $\overline{X}$; $\bar{x} = 62$ inches.
(*ii*) Sample proportion of households who liked particular programme P; $p = 0.85$.
(*iii*) Sample number of bolts according to specification X; $x = 20$, or sample proportion of bolts according to specification P; $p = 0.80$.
(*iv*) Sample mean consumption of electricity $\overline{X}$; $\bar{x} = 65$ units }

(*b*) Suppose I choose a random sample of three observations from a population and obtain the values 2, 5, 3. From these values I estimate the centre of the population by ranking the observations and taking the middle one. What estimator am I using and what is my estimate?

{ Sample median $X_{0.5} = X_{((n+1)/2)}$; $x_{0.5} = x_{((n+1)/2)} = 3$ }

3. (*a*) Is an estimator a random variable? Why or why not?
(Yes. An estimator is a random variable having its own probability distribution.)

(*b*) Why we call the standard deviation of a sample statistic as standard error of the statistic.
{ In the context of estimation, the deviation of a sample statistic T from its target θ (parameter to be estimated) must be considered an error. So the standard deviation of a sample statistic is commonly called the standard error of the sample statistic. }

4. (*a*) What is meant by unbiasedness? Differentiate between an unbiased and a biased estimator.

(*b*) A finite population consists of the numbers 3, 5, 7 and 9. Take all possible samples of size 2 which can be drawn with replacement form this population. By forming the sampling distributions of $\overline{X}$ and S^2 show that

(*i*) the sample mean $\overline{X} = \sum X_i / n$ is an unbiased estimator of the population mean μ.

(*ii*) the sample variance $S^2 = \sum (X_i - \overline{X})^2 / n$ is a biased estimator of the population variance σ^2.

{ (*i*) $\mu = 6$, $E(\overline{X}) = 6$, $E(\overline{X}) = \mu$, (*ii*) $\sigma^2 = 5$, $E(S^2) = 2.5$, $E(S^2) \neq \sigma^2$ }

(*c*) Draw all possible samples of size 3 taken without replacement from the population 7, 10, 13 and 16. By forming the sampling distributions, show that both sample mean $\overline{X}$ and sample median $X_{0.5}$ are unbiased estimators.

5. (*a*) A finite population consists of the numbers 1, 3, 5, 7 and 9. Consider all possible samples of size two which can be drawn with replacement form this population. By forming the sampling distributions of $\overline{X}$, S^2 and $\hat{S}^2$ show that

(*i*) the sample mean $\overline{X} = \sum X_i / n$ is an unbiased estimator of the population mean μ.

(*ii*) the sample variance $S^2 = \sum (X_i - \overline{X})^2 / n$ is a biased estimator of the population variance σ^2.

(*iii*) the sample variance $\hat{S}^2 = \sum (X_i - \overline{X})^2 / (n-1)$ is an unbiased estimator of the population variance σ^2.

{ (*i*) $\mu = 5$, $E(\overline{X}) = 5$, $E(\overline{X}) = \mu$, (*ii*) $\sigma^2 = 8$, $E(S^2) = 4$, $E(S^2) \neq \sigma^2$
(*iii*) $\sigma^2 = 8$, $E(\hat{S}^2) = 8$, $E(\hat{S}^2) = \sigma^2$ }

(*b*) A finite population consists of the numbers 2, 3, 4, 5, 6 and 8. Find the proportion P of even numbers in all possible random samples of size $n = 2$ that can be drawn with replacement from this population. By forming the sampling distribution of sample proportions show that sample proportion is an unbiased estimator of the population proportion. Also verify the relation

$$Var(P) = \frac{\pi(1-\pi)}{n}$$

where P and π are sample and population proportions respectively.

$\{ \pi = 2/3, \mu_P = 2/3, \sigma_P^2 = 1/9 \}$

(*c*) A finite population consists of the numbers 4, 5, 6 and 8. Find the proportion P of even numbers in all possible random samples of size $n = 3$ that can be drawn without replacement from this population. By forming the sampling distribution of sample proportions show that sample proportion is an unbiased estimator of the population proportion. Also verify the relation

$$Var(P) = \frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right)$$

where P and π are sample and population proportions respectively.

$\{ \pi = 3/4, \mu_P = 3/4, \sigma_P^2 = 1/48 \}$

◆

12.7 INTERVAL ESTIMATION

Interval estimation is a procedure of constructing an interval from a random sample, such that prior to sampling, it has a high specified probability of including the unknown true value of a population parameter.

12.7.1 Need for Interval Estimation. Any point estimate has the limitation that it does not provide information about the precision of the estimate *i. e.*, about the magnitude of error due to sampling. Often such information is essential for proper interpretation of the sample result.

A point estimator, calculated from the sample data, provides a single number as an estimate of the parameter. This single number lies in the fore front even though a statement of accuracy in terms of the standard error is attached to it. A point estimator, however efficient it may be, cannot be expected to be exactly equal to the population parameter. Moreover, we cannot assess simply by looking at just only one value (point estimator) how close is the estimate to the unknown true value of the parameter being estimated. A point estimate by itself does not supply this information about its precision.

An alternative approach to estimation is to extend the concept of error bound to produce an interval of values that is likely to include the unknown true value of the parameter. This is the concept underlying estimation by confidence intervals.

12.7.2 Interval Estimate. An *interval estimate* is an interval calculated from a random sample, such that prior to sampling, it has a high specified probability of including the unknown true value of a population parameter.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a population with unknown parameter θ. A confidence interval for θ is an interval (L , U) computed from the sample observations $X_1, X_2, \cdots, X_n$, such that prior to sampling, it includes the unknown true value of θ with a specified high probability. Let $(1 - \alpha)$ be a specified high probability and L and U be functions of sample observations $X_1, X_2, \cdots, X_n$ such that

$$P(L < \theta < U) = 1 - \alpha, \qquad \text{for } 0 < \alpha < 1$$

Then the interval (L, U) is called a $100(1-\alpha)\%$ *confidence interval* for the parameter θ, and the probability $(1-\alpha)$ is called the *confidence coefficient* or the *level of confidence*. Note that, $(1-\alpha)$ is the probability that the random interval (L, U) includes the parameter θ and not the probability that θ lies in the interval (L, U). The end points L and U that bound the confidence interval, are called the *lower* and *upper confidence limits* for the parameter θ. These limits being the functions of sample observations are random variables. The width $U - L$ of the confidence interval measures the precision of the estimate. The shorter the confidence interval, the more precise the estimate will be. The precision can be increased by

(*i*) decreasing the standard error of the estimate (*i. e.*, increasing the sample size).

(*ii*) decreasing the confidence coefficient.

12.7.3 Confidence Coefficient.

Meaning of Confidence Coefficient. From the definition of a confidence interval, we know that, prior to selecting the random sample, the probability is $1 - \alpha$ that the confidence interval we obtain will include the population parameter θ. The particular confidence interval result will be either correct or incorrect, and we do not know for certain which is the case.

Selecting the Confidence Coefficient. We should like the confidence interval to be very precise (*i. e.*, very narrow) and would like to be very confident that it includes θ. Unfortunately, for any fixed sample size, the confidence coefficient can only be increased by increasing the width of the confidence interval. The confidence interval widens rapidly as the confidence coefficient gets near 100 percent.

The choice of $1 - \alpha$ will vary from case to case, depending on how much risk of obtaining an incorrect interval can be taken. The numerical confidence coefficient (*e. g.*, 0.95) is often expressed as a percent (*e. g.*, 95%). Confidence coefficients of 90, 95, 98, and 99 percent are often used in practice.

12.8 CONFIDENCE INTERVAL FOR POPULATION MEAN, μ

The interval (L, U) is a $100(1-\alpha)\%$ confidence interval for the population mean μ if prior to sampling:

$$P(L < \mu < U) = 1 - \alpha$$

This definition simply states that a confidence interval with confidence coefficient $1 - \alpha$ is an interval estimate such that the probability is $1 - \alpha$ that the calculated limits include μ for any random sample. In other words, in many repeated random samples of size n from a population, $100(1-\alpha)\%$ of the interval estimates will include μ and therefore will be correct and $\alpha\%$ of the interval estimates will not include μ and therefore will be incorrect. The choice of method used in constructing a confidence interval for μ depends upon whether or not the population is normal, whether the population variance σ^2 is known or unknown, and whether the sample size n is large or small. We discuss these different cases below.

12.8.1 Normal Population, σ^2 known. Suppose that a random sample $X_1, X_2, \cdots, X_n$ of size n is drawn from a normal population with unknown mean μ and known variance σ^2. We wish to construct a confidence interval which is likely to include the true unknown value of the population mean μ with a degree of confidence $1 - \alpha$. We know that the sampling distribution

of $\bar{X}$ will be normal with mean μ and variance σ^2/n. Consequently, the distribution of the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

will be normal with mean 0 and variance 1. Then a two-sided $100(1 - \alpha)\%$ confidence interval for population mean μ is given by

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If $\bar{x}$ is the mean of an observed random sample of size n taken from a normal population with unknown mean μ and known variance σ^2, then a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

This can be written $\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

Note that, often the word *central* is omitted when considering confidence intervals, but it is assumed that a two-sided interval that is central, or symmetric about the mean is required.

12.8.2 Interpretation of a Confidence Interval. A $100(1 - \alpha)\%$ confidence interval for μ is

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If we identify

$$L = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U = \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

the probability statement implies that prior to sampling, the random interval (L, U) will include the parameter μ with a probability $1 - \alpha$. That is

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

It is to be emphasized that in this expression, μ is constant and it is the end points which are random variables. To better understand the meaning of a confidence statement, we perform repeated samplings from a normal distribution with mean μ and standard deviation σ and a $100(1 - \alpha)\%$ confidence interval $\bar{x} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ is computed from each random sample, approximately $100(1 - \alpha)\%$ of the intervals derived would contain the true value of μ.

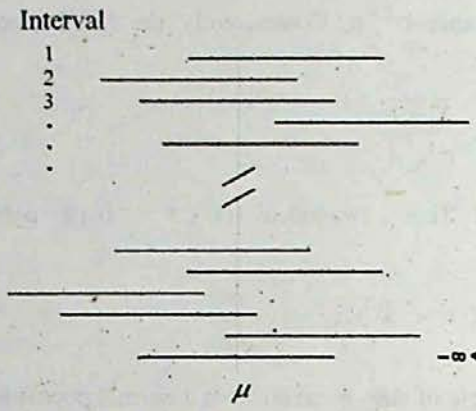


Fig. 12.1 Repeated forming confidence intervals for μ

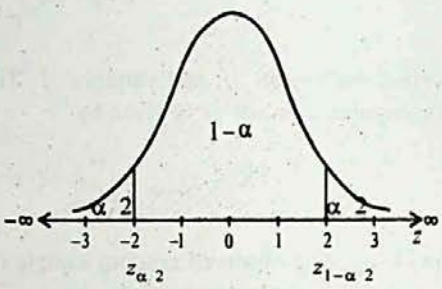


Fig. 12.2 Two-sided confidence interval for μ

Figure 12.1 shows what would typically happen if a number of samples were drawn from the same population and a confidence interval for μ were computed for each sample. The true value of μ is indicated by a vertical line in the figure. Different confidence intervals, resulting from different random samples, are shown as horizontal line segments. Most of the confidence intervals would contain μ, but some of them would not contain μ. If a 95% confidence interval were calculated for each sample then in the long-run, 95% of the confidence intervals that were formed would contain μ. That is not surprising, because the specified probability 0.95 represents the long-run relative frequency of these intervals crossing the vertical line.

Figure 12.2 represents that, before the sample is taken, the probability is $1 - \alpha$ that the quantity $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ will fall in the shaded interval. The interval estimate $\bar{X} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ will be correct (*i. e.*, will include μ) if $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ does fall in the shaded interval. In effect, the risk α of an incorrect confidence interval is divided equally in the two tails of the standard normal distribution.

12.8.3 Steps to Follow When Forming a Confidence Interval. We will follow the following standard format when estimating parameters with confidence intervals:

(*i*) Identify the population of interest, and state the conditions required for the validity of the procedure being used to construct the confidence interval.

(*ii*) Give the procedure (formula) that will be used

(*iii*) Construct the confidence interval

(*iv*) Interpret the results.

***Example* 12.7** *A normal population has a variance of* 100. *A random sample of size* 16 *selected from the population has a mean of* 52.5. *Construct the* 90% *confidence interval estimate of the population mean,* μ. *Interpret the result.*

Solution. The size and mean of sample and the variance of normal population are

$$n = 16, \qquad \bar{x} = 52.5, \qquad \sigma^2 = 100 \quad \Rightarrow \quad \sigma = 10,$$

Confidence coefficient: $1 - \alpha = 0.90$

$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$

$z_{1-\alpha/2} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

The two-sided 90% confidence interval for μ is

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$52.5 - (1.645)\frac{10}{\sqrt{16}} < \mu < 52.5 + (1.645)\frac{10}{\sqrt{16}}$$

$$48.4 < \mu < 56.6$$

A two-sided 90% confidence interval for μ obtained from the observed sample is (48.4, 56.6). We are 90% confident that the interval estimate contains μ.

***Example* 12.8** *Unoccupied seats on flights cause the airlines to lose revenue. Suppose a large airline obtained the 90% confidence interval for the average number of unoccupied seats per flight, on the basis of the records of its randomly selected 225 flights over the past year, as 11.15 to 12.05. Find the value of $\bar{x}$, the mean of the sample and σ the standard deviation of the normal population from which the sample was drawn. Estimate the average number of unoccupied seats per flight over the past year with 99% confidence coefficient.*

Solution. Sample size $n = 225$

Confidence coefficient: $1 - \alpha = 0.90$

$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$

$z_{1-\alpha/2} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

The two-sided $100(1 - \alpha)\%$ confidence interval for μ is

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The lower and upper limits of 90% confidence interval for μ are 11.15 and 12.05. Thus

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 11.15 \Rightarrow \bar{x} - (1.645)\frac{\sigma}{\sqrt{225}} = 11.15 \quad \ldots\ldots(i)$$

$$\bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 12.05 \Rightarrow \bar{x} + (1.645)\frac{\sigma}{\sqrt{225}} = 12.05 \quad \ldots\ldots(ii)$$

Adding (*i*) and (*ii*), we get

$2\bar{x} = 23.20 \Rightarrow \bar{x} = 11.6$

Putting the value of $\bar{x}$ in (*ii*), we get

$$11.6 + (1.645)\frac{\sigma}{\sqrt{225}} = 12.05 \Rightarrow \sigma = 4.1$$

Confidence coefficient: $1 - \alpha = 0.90$

$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow 1 - \alpha/2 = 0.995$

$z_{1-\alpha/2} = z_{0.995} = 2.576$ { From Table 10 (*b*) }

The two-sided 90% confidence interval for μ is

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$11.6 - (2.576)\frac{4.1}{\sqrt{225}} < \mu < 11.6 + (2.576)\frac{4.1}{\sqrt{225}}$$

$$10.9 < \mu < 12.3$$

12.8.4 Any Population, σ^2 known/unknown, n large. Suppose that a random sample $X_1, X_2, \cdots, X_n$ of size n is drawn from a population with mean μ and variance σ^2. We wish to construct a confidence interval which is likely to trap the true unknown value of the population mean μ with a degree of confidence $1 - \alpha$. If the population is not normal, and if σ^2 is either known or unknown, then according to the Central Limit Theorem the sampling distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n (when σ^2 is known and $\hat{S}^2/n$ when σ^2 is unknown) if the sample size is sufficiently large, say, $n > 30$. Consequently the distribution of the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cong \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}}$$

is approximately normal with mean 0 and variance 1. Then a two sided $100(1 - \alpha)\%$ approximate confidence interval for population mean μ is given by

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We now turn to the more realistic situation for which the population variance σ^2 is unknown. Because n is large, replacing σ with its best unbiased estimator $\hat{S}$ does not appreciably affect the probability statement. When n is large and population variance σ^2 is unknown, a $100(1 - \alpha)\%$ approximate confidence interval for population mean μ is given by

$$\bar{X} - z_{1-\alpha/2}\frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\hat{S}}{\sqrt{n}}.$$

If $\bar{x}$ and $\hat{s}$ is the mean and standard deviation of an observed random sample of size n sufficiently large from a population with unknown mean μ and unknown but finite variance σ^2, then a $100(1 - \alpha)\%$ approximate confidence interval for μ is given by

$$\bar{x} - z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

This can be written $\bar{x} \pm z_{1-\alpha/2} \dfrac{\hat{s}}{\sqrt{n}}$

12.8.5 Sampling Without Replacement. When sampling is done without replacement from a finite population of size N, the standard error of $\bar{X}$ is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

If the sample size n is greater than 5% of the population size N (*i. e.*, $n > 0.05\,N$), then a $100\,(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

However, when n is large and population variance σ^2 is unknown, a $100\,(1-\alpha)\%$ approximate confidence interval for population mean μ is given by

$$\bar{X} \pm z_{1-\alpha/2} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

If $\bar{x}$ is the mean of an observed random sample of size n taken from a population with unknown mean μ and known variance σ^2, then a $100\,(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

If $\bar{x}$ and $\hat{s}$ is the mean and standard deviation of an observed random sample of size n sufficiently large from a population with unknown mean μ and unknown but finite variance σ^2, then a $100\,(1-\alpha)\%$ approximate confidence interval for μ is given by

$$\bar{x} \pm z_{1-\alpha/2} \frac{\hat{s}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

The finite population correction $(N-n)/(N-1)$ may be ignored when the sample size n is less than 5% of the population size N (*i. e.*, $n < 0.05\,N$).

***Example* 12.9** *A particular component in a transistor circuit has a lifetime which is known to follow a skew distribution. A random sample of* 250 *components from a week's production given an average lifetime of* 840 *hours, and the variance of lifetimes is* 483 (*hours*^{2}). *Find approximately* 95% *confidence limits to the true mean lifetime in the whole population of the product.*

Solution. The size, mean and variance of the sample are

$$n = 250, \quad \bar{x} = 840, \quad \hat{s}^2 = 483 \Rightarrow \hat{s} = \sqrt{483} = 21.98$$

Confidence coefficient : $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{ \text{From Table 10 } (b) \}$$

The two-sided 95% approximate confidence interval for μ is

$$\bar{x} - z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

$$840 - (1.960)\frac{21.98}{\sqrt{250}} < \mu < 840 + (1.960)\frac{21.98}{\sqrt{250}}$$

$$837.3 < \mu < 842.7$$

***Example* 12.10** *A random sample of size* $n = 200$, *selected without replacement from a finite population of size* $N = 1000$ *with* $\sigma = 1.28$, *showed that* $\bar{x} = 68.6$. *Construct a* 97% *confidence interval for the mean* μ *of the population.*

Solution. The size and mean of sample and the size and a standard deviation of population are

$$n = 200, \quad \bar{x} = 68.6, \quad N = 1000, \quad \sigma = 1.28$$

Confidence coefficient: $1 - \alpha = 0.97$

$$1 - \alpha = 0.97 \Rightarrow \alpha = 0.03 \Rightarrow \alpha/2 = 0.015 \Rightarrow 1 - \alpha/2 = 0.985$$

$$z_{1-\alpha/2} = z_{0.985} = 2.17 \qquad \{ \text{From Table 10 } (a) \}$$

The two-sided 97% approximate confidence interval for μ is

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$68.6 \pm (2.17)\frac{1.28}{\sqrt{200}}\sqrt{\frac{1000-200}{1000-1}}$$

$$(68.42, 68.78) \qquad \Rightarrow \qquad 68.42 < \mu < 68.78$$

***Example* 12.11** *An auditor has selected a simple random sample of* 100 *accounts from the* 8042 *accounts receivable of a freight company to estimate the total audit amount of the receivable in the population. The sample mean is* $\bar{x} = 33.19$ *and the sample standard deviation is* $\hat{s} = 34.48$. *Obtain the* 95.44 *percent confidence interval for the mean audit amount in the population.*

Solution. The size, mean and standard deviation of the sample and the population size are

$$n = 100, \quad \bar{x} = 33.19, \quad \hat{s} = 34.48, \quad N = 8042$$

Confidence coefficient: $1 - \alpha = 0.9544$

$$1 - \alpha = 0.9544 \Rightarrow \alpha = 0.0456 \Rightarrow \alpha/2 = 0.0228 \Rightarrow 1 - \alpha/2 = 0.9772$$

$$z_{1-\alpha/2} = z_{0.9772} = 2 \qquad \{ \text{From Table 10 } (a) \}$$

The two-sided 95.44% approximate confidence interval for the mean audit amount μ is

$$\bar{x} \pm z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$33.19 \pm (2)\frac{34.48}{\sqrt{100}}\sqrt{\frac{8042-100}{8042-1}}$$

$$(26.3366, 40.0434) \qquad \Rightarrow \qquad 26.3366 < \mu < 40.0434$$

12.8.6 Normal Population, σ^2 unknown, n small. Suppose that a random sample $X_1, X_2, \cdots, X_n$ of size n is drawn from a normal population with unknown mean μ and unknown variance σ^2. We wish to construct a confidence interval which is likely to contain the true unknown value of population mean μ with a confidence coefficient $1 - \alpha$. However, time and cost restrictions would probably limit the sample size to a small number. Many inferences in business must be made on the basis of very limited information, *i. e., small samples.* When the population is normal, the sampling distribution of the statistic

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

is a t-distribution with $\nu = n - 1$ degrees of freedom. Thus, when n is small, population is normal and population variance is unknown, a two-sided $100(1 - \alpha)\%$ confidence interval for population mean μ is given by

$$\overline{X} - t_{\nu;1-\alpha/2}\frac{\hat{S}}{\sqrt{n}} < \mu < \overline{X} + t_{\nu;1-\alpha/2}\frac{\hat{S}}{\sqrt{n}}$$

If $\bar{x}$ and $\hat{s}$ is the mean and standard deviation of an observed random sample of size n from a normal population with unknown mean μ and unknown but finite variance σ^2, then a $100(1 - \alpha)\%$ interval for μ is given by

$$\bar{x} - t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

This can be written $$\bar{x} \pm t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

For ***large degrees of freedom*** (e.g., beyond the range of Table 12) the ***t-distribution*** can be approximated by a ***standard normal distribution.***

***Example* 12.12** *Ten packets of a particular brand of biscuits are chosen at random and their mass measured in grams. The results are*

$$n = 10, \qquad \sum x_i = 3978.7, \qquad \sum x_i^2 = 1583098.3$$

Assuming that the sample is taken from a normal population with mean mass μ, calculate the 98% confidence interval for μ.

Solution. The mean and standard deviation of the sample are

$$\bar{x} = \frac{\sum x}{n} = \frac{3978.7}{10} = 397.87$$

$$\hat{s} = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{1583098.3 - 10(397.87)^2}{10-1}} = 3.213$$

Confidence coefficient: $1 - \alpha = 0.98$

$$1 - \alpha = 0.98 \Rightarrow \alpha = 0.02 \Rightarrow \alpha/2 = 0.01 \Rightarrow 1 - \alpha/2 = 0.99$$

Degrees of freedom : $\nu = n - 1 = 10 - 1 = 9$

$t_{\nu;1-\alpha/2} = t_{9;0.99} = 2.821$ (From Table 12)

The two-sided 98% confidence interval for μ is

$$\bar{x} - t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

$$397.87 - (2.821)\frac{3.213}{\sqrt{10}} < \mu < 397.87 + (2.821)\frac{3.213}{\sqrt{10}}$$

$$395.004 < \mu < 400.736$$

Exercise 12.2

1. (*a*) What is meant by estimation ? Distinguish between point estimate and interval estimate. Why is an interval estimate more useful ?

(*b*) Distinguish between the following

(*i*) Estimator and estimate,

(*ii*) Point and interval estimation..

2. (*a*) Explain what is meant by

(*i*) Confidence interval,

(*ii*) Confidence limits,

(*iii*) Confidence coefficient.

(*b*) When would a confidence interval be preferred over point estimation for a parameter. (When the reliability of a point estimator is needed, the confidence interval conveniently express the estimator along with its measure of variation. Reliability is reported through the confidence coefficient and the variation is reflected in the length of the interval.)

3. (*a*) Find a 90% confidence interval for the mean of a normal distribution with $\sigma = 3$, given the sample as 2.3, −0.2, −0.4, −0.9.
($-2.268 < \mu < 2.668$)

(*b*) The standard deviation of the amounts poured into bottles by an automatic filling machine is 1.8 *ml* (*millilitre*). The amounts of fill in a random sample of bottles, in *ml*, were 481, 479, 482, 480, 477, 478, 481 and 482. Suppose the population of amounts of fill is normal. Construct a 90% confidence interval for the mean amount in all bottles filled by the machine.
($478.95 < \mu < 481.05$)

4. (*a*) A random sample of size 36 is taken from a normal population with a known variance $\sigma^2 = 25$. If the mean of the sample is 42.6, find 95% confidence limits for the population mean.
($40.967 < \mu < 44.233$)

(*b*) A school wishes to estimate the average weight of students in the sixth grade. A random sample of $n = 25$ is selected and the sample mean is found to be $\bar{x} = 100$ *lbs*. The

standard deviation of the population is known to be 15 *lbs*. Compute 90% confidence interval for the population mean.
($95.065 < \mu < 104.935$)

5. (*a*) Suppose that the weights of 100 male students of a university represent a random sample of weights of 1546 students of the university. Find 99% confidence intervals for the mean weight of the students, given $\bar{x} = 67.45$ and $\hat{s} = 2.93$.
($66.68 < \mu < 68.22$)

(*b*) 150 bags of flour of a particular brand are weighed and the mean mass is found to be 748 *g* with standard deviation 3.6 *g*. Find 98% confidence intervals for the mean mass of bags of flour of this brand.
($747.316 < \mu < 748.684$)

6. (*a*) If the two-sided $100(1 - \alpha)\%$ confidence interval based on random sample taken from $X \sim N(\mu, \sigma^2)$ is $12.18 < \mu < 20.56$, find $\bar{x}$.
($\bar{x} = 16.37$)

(*b*) On the basis of the results obtained from a random sample of 100 men from a particular area, the 95% confidence interval for the mean height of the male population of the area was found to be (177.22 *cm*. to 179.18 *cm*). Find the value of $\bar{x}$, the mean of the sample and σ, the standard deviation of the normal population from which the sample was drawn. Find 98% confidence interval for the mean height.
($\bar{x} = 178.2$, $\sigma = 5$, $177.04 < \mu < 179.36$)

7. (*a*) Explain what is meant by the statement, 'we are $100(1 - \alpha)\%$ confident that our interval estimate contains μ.'
{ In repeated sampling, $100(1 - \alpha)\%$ of all such confidence intervals contain μ. }

(*b*) Explain what is meant by the statement, "we are 95% confident that our interval contains μ".
(In repeated sampling, 95% of all such confidence intervals contain μ.)

(*c*) If an 85% confidence interval is $27.5 < \mu < 43.8$, what does this statement mean ?
(Intervals so formed would contain μ 85% of the time.)

(*d*) If $\alpha = 0.10$, how many intervals would be expected to contain μ ?
(We would expect about 90% of the intervals to contain μ and 10% to miss μ in the long-run in repeated sampling.)

8. (*a*) What role does the sample mean play in a two-sided confidence interval for μ, based on a random sample from $X \sim N(\mu, \sigma^2)$?
(The sample mean is the midpoint of the confidence interval but has no effect on the length of the interval.)

(*b*) When setting a two-sided $100(1 - \alpha)\%$ confidence interval for μ, based on a random sample of size n from a normal population, how the following changes will affect the length of the confidence interval for μ: (Assume all other quantities remain fixed.)

(*i*)	increasing n	(*ii*)	increasing $(1 - \alpha)$
(*iii*)	decreasing n	(*iv*)	decreasing $(1 - \alpha)$
(*v*)	increasing $\hat{s}^2$	(*vi*)	increasing $\bar{x}$
(*vii*)	increasing α	(*viii*)	decreasing $\hat{s}$

(decreased, increased, increased, decreased, increased, no effect, decreased, decreased)

9. (*a*) Define Student's *t*-statistic. What assumptions are made about the population where the *t*-distribution is used?

(*b*) The contents of 10 similar containers of a commercial soap are : 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 and 9.8 litres. Find 99% confidence interval for the mean soap content of all such containers, assuming an approximate normal distribution. ($9.807 < \mu < 10.313$)

10. (*a*) The masses in grams, of thirteen ball bearings taken at random from a batch are 21.4, 23.1, 25.9, 24.7, 23.4, 21.5, 25.0, 22.5, 26.9, 26.4, 25.8, 23.2, 21.9. Calculate a 95% confidence interval for the mean mass of the population, supposed normal, from which these masses were drawn.
($22.82 < \mu < 25.14$)

(*b*) A random sample of seven independent observations of a normal variable gave $\sum x = 35.9$, $\sum x^2 = 186.19$. Calculate a 90% confidence interval for the population mean.
($4.70 < \mu < 5.56$)

11. (*a*) A random sample of eight observations of a normal variable gave $\sum x = 261.2$, $\sum (x - \bar{x})^2 = 3.22$. Calculate a 95% confidence interval for the population mean. ($32.08 < \mu < 33.22$)

(*b*) A sample of 12 measurements of the breaking strength of cotton threads gave a mean $\bar{x} = 209$ grams and a standard deviation $\hat{s} = 35$ grams. Find 95% and 99% confidence limits for the actual mean breaking strength.
($186.76 < \mu < 231.24$; $177.62 < \mu < 240.38$)

12. (*a*) A random sample of 16 values from a normal population showed a mean of 41.5 inches and a sum of squares of deviations from this mean equal to 135 $(\text{inches})^2$. Show that the 95% confidence limits for this mean are 39.9 and 43.1 inches.

(*b*) Find a 99% confidence interval for the mean of normal distribution with $\sigma = 2.5$ and if a sample of size 7 gave the values 9, 16, 10, 14, 8, 13, 14. What would be the confidence interval if σ were unknown.
($9.566 < \mu < 14.434$; $7.797 < \mu < 16.203$ when σ is unknown.)

◆

12.9 CONFIDENCE INTERVAL FOR POPULATION PROPORTION OF SUCCESSES, π

The interval (L, U) is a $100(1 - \alpha)\%$ confidence interval for the population proportion of successes π if prior to sampling

$$P(L < \pi < U) = 1 - \alpha$$

This definition simply states that a confidence interval with confidence coefficient $1 - \alpha$ is an interval estimate such that the probability is $1 - \alpha$ that the calculated limits include π for any random sampling. In other words, in many repeated random samples of size n from a Bernoulli population, $100(1 - \alpha)\%$ of the interval estimates will include the true population proportion

of successes π and therefore will be correct and $100\,\alpha\,\%$ of the interval estimates will not include π and therefore will be incorrect.

In many problems we must estimate the population proportion or percentage, for example, the proportion of defectives found in shipment of raw materials upon inspection. In this case it seems to be reasonable that we are sampling from a Bernoulli population; hence our problem is to estimate its parameter π. The interval is based on the estimator $P = X/n$, the sample fraction of successes. We know that the sampling distribution of P is a binomial distribution. The binomial distribution of the estimator P can be approximated by the normal distribution with a mean of $\mu_P = \pi$ and a standard deviation of $\sigma_P = \sqrt{\pi(1-\pi)/n}$, when n is large and π is not too near 0 or 1. Consequently the distribution of the statistic

$$Z = \frac{P - \pi}{\sqrt{P(1-P)/n}}$$

will be approximately normal with mean 0 and variance 1. Then a two-sided confidence interval for population proportion of successes π is given by

$$P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} < \pi < P + z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

If $p = x/n$ is the proportion of successes in an observed random sample of size n, then a $100\,(1-\alpha)\%$ confidence interval for π is given by

$$p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

This can be written $\quad p \pm z_{1-\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$

12.9.1 Sampling Without Replacement. When sampling is done without replacement from a finite population of size N, the standard error of P is given by

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

which is estimated as

$$\hat{\sigma}_P = \sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}$$

If the sample size n is greater than 5% of the population size N (*i. e.*, $n > 0.05\,N$), then a $100\,(1-\alpha)\%$ confidence interval for π is given

$$P \pm z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}$$

The finite population correction $(N-n)/(N-1)$ may be ignored when the sample size n is less than 5% of the population size N (*i. e.*, $n < 0.05\,N$).

***Example* 12.13** *In a random sample of* 500 *young persons from a small town* 40 *were found to be unemployed. Compute a* 96% *confidence interval for the rate of unemployment in the town. Interpret the result.*

Solution. The sample size, number of successes and proportion of successes in the sample are

$$n = 500, \quad x = 40, \quad p = \frac{x}{n} = \frac{40}{500} = 0.08$$

Confidence coefficient : $1 - \alpha = 0.96$

$$1 - \alpha = 0.96 \Rightarrow \alpha = 0.04 \Rightarrow \alpha/2 = 0.02 \Rightarrow 1 - \alpha/2 = 0.98$$

$$z_{1-\alpha/2} = z_{0.98} = 2.054 \qquad \{ \text{From Table 10 } (a) \}$$

The two-sided 96% confidence interval for π is

$$p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$0.08 - 2.054\sqrt{\frac{0.08(1-0.08)}{500}} < \pi < 0.08 + 2.054\sqrt{\frac{0.08(1-0.08)}{500}}$$

$$0.055 < \pi < 0.105$$

We are 96% confident that rate of unemployment is between 5.5% to 10.5% because our procedure will produce true statement 96% of the time.

***Example* 12.14** *A poll is taken among the residents of a city and the surrounding country to determine the feasibility of a proposal to construct a civic centre. If* 2400 *of* 5000 *city residents favour it, find almost certain limits for the true fraction favouring the proposal to construct the civic centre.*

Solution. The sample size, number of successes and proportion of successes in the sample are

$$n = 5000, \quad x = 2400, \quad p = \frac{x}{n} = \frac{2400}{5000} = 0.48$$

Confidence coefficient : $1 - \alpha = 0.999$ (almost certain is 99.9% confident)

$$1 - \alpha = 0.999 \Rightarrow \alpha = 0.001 \Rightarrow \alpha/2 = 0.0005 \Rightarrow 1 - \alpha/2 = 0.9995$$

$$z_{1-\alpha/2} = z_{0.9995} = 3.291 \qquad \{ \text{From Table 10 } (b) \}$$

The two-sided 99.9% confidence interval for π is

$$p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$0.48 - 3.291\sqrt{\frac{0.48(1-0.48)}{5000}} < \pi < 0.48 + 3.291\sqrt{\frac{0.48(1-0.48)}{5000}}$$

$$0.457 < \pi < 0.503$$

We are almost certain that the true fraction favouring the proposal to construct the civic centre lies between 0.457 and 0.503.

***Example* 12.15** *A random sample of* 250 *from the* 5000 *students in Govt. College, Gujranwala contained* 30 *left-handed students. Give an approximate* 95% *confidence interval for the proportion of left-handed students in the college.*

Solution. The sample size, number of successes and proportion of successes in the sample and the population size are

$$n = 250, \quad x = 30, \quad p = \frac{x}{n} = \frac{30}{250} = 0.12, \quad N = 5000$$

Confidence coefficient : $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{ \text{From Table } 10\,(b) \}$$

The two-sided 95% confidence interval for π in the finite population is

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$0.12 \pm 1.960 \sqrt{\frac{0.12(1-0.12)}{250}} \sqrt{\frac{5000-250}{5000-1}}$$

$$(0.081, \ 0.159) \qquad \Rightarrow \qquad 0.081 < \pi < 0.159$$

Exercise 12.3

1. (*a*) A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favour of a particular candidate. Find (*i*) 95% and (*ii*) 99% confidence limits for the proportion of all the voters in favour of this candidate.

{ (*i*) $0.453 < \pi < 0.647$, (*ii*) $0.422 < \pi < 0.678$ }

(*b*) In a random sample of 1000 houses in a certain city, it is found that 228 own colour television sets. Find 98% confidence interval for the proportion of houses in this city that have coloured sets.

$(0.197 < \pi < 0.259)$

2. (*a*) In 40 tosses of a coin 24 heads were obtained. Find (*i*) 95% and (*ii*) 99.73% confidence limits for the proportion of heads which would be obtained in an unlimited number of tosses of the coin.

{ (*i*) $0.448 < \pi < 0.752$, (*ii*) $0.368 < \pi < 0.832$ }

(*b*) A random sample of 200 voters in a constituency included 110 who said they would vote for Mr. *A*. Assuming all the 15000 voters in the constituency would vote, give an approximate 95% confidence interval for the proportion who would vote for Mr. *A*.

$(0.4815 < \pi < 0.6185)$

(*c*) A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the percentage of bad pineapples in the consignment almost certainly lies between 8.05 and 17.95.

◆

12.10 COMPARATIVE STUDIES

To this point we have been concerned with inferences about parameters of a single population. We now turn our attention to estimation procedures that are important in comparing the parameter values of two populations. To make inferences about two populations; we must obtain two samples — one from each population. There are many methods by which the two samples could be obtained; we will discuss two of them in this text. These methods result in either *independent* or *dependent* samples.

12.10.1 Independent Samples. If two samples are selected, one from each of two populations, then the two samples are *independent* if the selection of objects from one population is unrelated to the selection of objects from the other population.

12.10.2 Dependent Samples. If two samples are selected, one from each of two populations then the two samples are *dependent*, if for each object selected from one population an object is chosen from the other population to form a pair of similar objects. These samples are also called as matched samples. The set of sample pairs is called a paired samples.

The key to recognizing two independent samples is to realize that they are always two different random samples, whereas the dependent samples always consist of matched, or paired, observations.

12.11 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO POPULATION MEANS, $\mu_1 - \mu_2$

In many business and management problems, we wish to estimate the difference between the means of two populations. For instance, we, may want to decide upon the basis of suitable samples to what extent, if any, a fertilizer is more effective than an existing fertilizer; a newly introduced product is more reliable than an existing product, or the degree to which a particular training programme improves worker attitudes or performance.

12.11.1 Independent Samples: ***Normal populations, known variances, any sample size.*** If $\bar{X}_1$ and $\bar{X}_2$ are respectively, the means of independent random samples of sizes n_1 and n_2 taken from two normal populations having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, then $\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n)$ and $\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n)$ and that $\bar{X}_1$ is independent of $\bar{X}_2$.

Since any linear combination of independent normal random variables is also normally distributed, then $\bar{X}_1 - \bar{X}_2$ is a random variable having a normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Thus the distribution of random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is a standard normal distribution. Then a two-sided $100\,(1 - \alpha)\%$ confidence interval for difference between means of the two populations $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If $\bar{x}_1$ and $\bar{x}_2$ are the means of the two independent observed samples, then a $100\,(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

***Example* 12.16** *Apex's current packing machinery is known to pour ground coffee into "1-pound cans" with a standard deviation of* 0.6 *ounce. Apex is considering using a new packing machine which is said to pour coffee into "1-pound cans" more accurately, with a standard deviation of* 0.3 *ounce. Both machines pour ground coffee according to a normal distribution. Before deciding to invest, Apex wishes to evaluate the performance of the new machine against that of the old machine. A sample was taken on each machine against that of mean weight of the contents of the "1-pound cans" yielding the following result.*

Sample	Size	Mean
Using Old Machine: I	$n_1 = 25$	$\bar{x}_1 = 16.7$
Using New Machine: II	$n_2 = 36$	$\bar{x}_2 = 15.8$

Construct a 95% *confidence interval for the difference in the average weight of the contents poured by the old versus the new machine.*

Solution. The sizes and means of two samples and the standard deviations of two populations are

$$n_1 = 25, \qquad \bar{x}_1 = 16.7, \qquad \sigma_1 = 0.6$$

$$n_2 = 36, \qquad \bar{x}_2 = 15.8, \qquad \sigma_2 = 0.3$$

Confidence coefficient : $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{\text{From Table 10 } (b)\}$$

The two-sided 95% confidence limits for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(16.7 - 15.8) \pm 1.960\sqrt{\frac{(0.6)^2}{25} + \frac{(0.3)^2}{36}}$$

$$(0.65,\ 1.15) \qquad \Rightarrow \qquad 0.65 < \mu_1 - \mu_2 < 1.15$$

12.11.2 Independent Samples: ***Any populations, variances known/unknown, large samples.*** When both sample sizes are large (say greater than 30) the assumptions regarding small samples can be greatly relaxed. It is no longer necessary to assume that the parent distributions are normal, because the Central Limit Theorem assures that $\bar{X}_1$ is approximately normally distributed with mean μ_1 and variance σ_1^2/n_1, and that $\bar{X}_2$ is also approximately normally distributed with mean μ_2 and variance σ_2^2/n_2, and that $\bar{X}_1$ is independent of $\bar{X}_2$, then $(\bar{X}_1 - \bar{X}_2)$ is approximately normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

Thus the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has an approximately standard normal distribution. Because n_1 and n_2 are both large, the approximation remains valid of if σ_1^2 and σ_2^2 are replaced by their sample variances $\hat{S}_1^2$ and $\hat{S}_2^2$. The assumption of equal variance is not required in inferences derived from large samples. We can modify the previous result to obtain a confidence interval by substituting the sample variances $\hat{S}_1^2$ for σ_1^2 and $\hat{S}_2^2$ for σ_2^2 as long as both samples are large enough ($n_1 > 30$, $n_2 > 30$) for the Central Limit Theorem to be invoked. Hence the distribution of random variable $\bar{X}_1 - \bar{X}_2$ approaches a normal distribution with mean $\mu_1 - \mu_2$ and variance $\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2$.

Then the distribution of random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

approaches the standard normal distribution. Then a two-sided $100(1 - \alpha)\%$ approximate confidence interval for difference between means of the two populations $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

If $\bar{x}_1$ and $\bar{x}_2$ are the means and $\hat{s}_1^2$ and $\hat{s}_2^2$ are the variances of the two independent observed random samples, then a $100(1 - \alpha)\%$ approximate confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$$

***Example* 12.17** *Rural and urban students are to be compared on the basis of their scores on a nation wide musical aptitude test. Two random samples of sizes* 90 *and* 100 *are selected from rural and urban seventh class students. The summary statistics from the test scores are*

Sample	Size	Mean	Standard deviation
Rural: I	$n_1 = 90$	$\bar{x}_1 = 76.4$	$\hat{s}_1 = 8.2$
Urban: II	$n_2 = 100$	$\bar{x}_2 = 81.2$	$\hat{s}_2 = 7.6$

Establish a 98% *confidence interval for the difference in population mean scores between urban and rural students.*

Solution. *Confidence coefficient*: $1 - \alpha = 0.98$

$$1 - \alpha = 0.98 \Rightarrow \alpha = 0.02 \Rightarrow \alpha/2 = 0.0 \Rightarrow 1 - \alpha/2 = 0.99$$

$$z_{1-\alpha/2} = z_{0.99} = 2.326 \quad \{ \text{From Table 10 } (b) \}$$

The two-sided 98% approximate confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{x}_2 - \bar{x}_1) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$$

$$(81.2 - 76.4) \pm 2.326 \sqrt{\frac{(7.6)^2}{100} + \frac{(8.2)^2}{90}}$$

$$(2.1,\ 7.5) \quad \Rightarrow \quad 2.1 < \mu_2 - \mu_1 < 7.5$$

We conclude, with 98% confidence, that the mean of urban scores is at least 2.1 units higher and can be as much as 7.5 units higher than the mean of rural scores.

12.11.3 Independent Samples: ***Normal populations, same unknown variance, small samples.*** When n_1 and n_2 are small and σ_1^2 and σ_2^2 are unknown, the formula for constructing a confidence interval that we have been discussing cannot be used. However, for independent samples from two normal populations having the same unknown variance σ^2, we can develop a confidence interval for $\mu_1 - \mu_2$ as follows :

If $\bar{X}_1$ and $\bar{X}_2$ are respectively, the means of two independent random samples taken from populations which are $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, then $\bar{X}_1 \sim N(\mu_1, \sigma^2/n)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2/n)$ and that $\bar{X}_1$ is independent of $\bar{X}_2$.

Since any linear combination of independent normal random variables is also normally distributed, then $\bar{X}_1 - \bar{X}_2$ is normally distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Thus the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a standard normal distribution. Thus, if $\hat{S}_1^2$ and $\hat{S}_2^2$ are the two sample variances (both estimating the variance σ^2 common to both populations), the pooled (weighted arithmetic mean) estimator of σ^2, denoted by S_p^2, is

$$S_p^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2} = \frac{\Sigma(X_{i1} - \bar{X}_1)^2 + \Sigma(X_{i2} - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{(\Sigma X_{i1}^2 - n_1 \bar{X}_1^2) + (\Sigma X_{i2}^2 - n_2 \bar{X}_2^2)}{n_1 + n_2 - 2}$$

Then the random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. Then a two-sided $100(1 - \alpha)\%$ confidence interval for difference between means of the two populations $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\nu;1-\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If $\bar{x}_1$ and $\bar{x}_2$ are the means of the two observed random samples and s_p is the pooled estimate of the common standard deviation of the two normal populations, then a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

It should be noted that this is used under the conditions when the sample sizes are small (*i. e.*, $n_1 \leq 30$ and $n_2 \leq 30$). When both n_1 and n_2 are greater than 30, it is legitimate to use

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

as a good approximation.

***Example* 12.18** *Suppose you want to estimate the difference in annual operation costs for automobiles with rotary engines and those with standard engines. You find 8 owners of cars with rotary engines and 12 owners with standard engines, who have purchased their cars within the last two years and are willing to participate in the experiment. Each of the 20 owners keeps accurate records of the amount spent on operating his or her car (including gasoline, oil, repairs, etc.) for a 12 month period. All costs are recorded on a per 1000 mile basis to adjust for differences in mileage driven during the 12 month period. The results are summarized below:*

Sample	*Size*	*Mean*	*Standard deviation*
Rotary: I	$n_1 = 8$	$\bar{x}_1 = 56.96$	$\hat{s}_1 = 4.85$
Standard: II	$n_2 = 12$	$\bar{x}_2 = 52.73$	$\hat{s}_2 = 6.35$

Estimate the true difference $(\mu_1 - \mu_2)$ *between the mean operating cost per* 1000 *miles of cars with rotary and standard engines. Use a* 90% *confidence level.*

Solution. The pooled estimate of population common standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(8 - 1)(4.85)^2 + (12 - 1)(6.35)^2}{8 + 12 - 2}} = 5.813$$

Confidence coefficient: $1 - \alpha = 0.90$

$$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$$

Degrees of freedom: $\nu = n_1 + n_2 - 2 = 8 + 12 - 2 = 18$

$t_{\nu;1-\alpha/2} = t_{18;95} = 1.734$ (From Table 12)

The two-sided 90% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(56.96 - 52.73) \pm 1.734\,(5.813) \sqrt{\frac{1}{8} + \frac{1}{12}}$$

$$(-0.37,\ 8.83) \quad \Rightarrow \quad -0.37 < \mu_1 - \mu_2 < 8.83$$

***Example* 12.19** *Given two random samples from two independent normal populations, with*

Sample	Size	Mean	Sum of squares
I	$n_1 = 11$	$\bar{x}_1 = 75$	$\sum (x_{i1} - \bar{x}_1)^2 = 372.1$
II	$n_2 = 14$	$\bar{x}_2 = 60$	$\sum (x_{i2} - \bar{x}_2)^2 = 365.17$

Find a 99% *confidence interval for* $(\mu_1 - \mu_2)$. *Assume that population variances are equal.*

Solution. The pooled estimate of population common standard deviation is

$$s_p = \sqrt{\frac{\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{372.1 + 365.17}{11 + 14 - 2}} = 5.66$$

Confidence coefficient: $1 - \alpha = 0.99$

$$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow 1 - \alpha/2 = 0.995$$

Degrees of freedom: $\nu = n_1 + n_2 - 2 = 11 + 14 - 2 = 23$

$t_{\nu;1-\alpha/2} = t_{23;0.995} = 2.807$ (From Table 12)

The two-sided 99% confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(75 - 60) \pm 2.807\,(5.66) \sqrt{\frac{1}{11} + \frac{1}{14}}$$

$$(8.6,\ 21.4) \quad \Rightarrow \quad 8.6 < \mu_1 - \mu_2 < 21.4$$

***Example* 12.20** *A course in mathematics is taught to 10 students by the conventional class room method. A second group of 12 students was given the same course by means of programmed materials. At the end of the semester, the same examination was given to each group. Their scores are given below:*

Group I	70	66	76	77	73	72	68	74	75	69		
Group II	77	83	92	85	82	84	80	86	91	93	80	87

Compute a 90% confidence interval for the difference between the average scores of the two populations. Assume the populations to be approximately normal with equal variance

Solution. The sample means and pooled estimate of population common standard deviation are

x_1	70	66	76	77	73	72	68	74	75	69			$\Sigma x_1 = 720$
x_2	77	83	92	85	82	84	80	86	91	93	80	87	$\Sigma x_2 = 1020$
x_1^2	4900	4356	5776	5929	5329	5184	4624	5476	5625	4761			$\Sigma x_1^2 = 51960$
x_2^2	5929	6889	8464	7225	6724	7056	6400	7396	8281	8649	6400	7569	$\Sigma x_2^2 = 86982$

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{720}{10} = 72$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{1020}{12} = 85$$

$$s_p = \sqrt{\frac{(\Sigma x_1^2 - n_1 \bar{x}_1^2) + (\Sigma x_2^2 - n_2 \bar{x}_2^2)}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{\{51960 - 10(72)^2 + \{86982 - 12(85)^2\}}{10 + 12 - 2}} = 4.48$$

Confidence coefficient : $1 - \alpha = 0.90$

$$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$$

Degrees of freedom : $\nu = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$

$$t_{\nu;1-\alpha/2} = t_{20;0.95} = 1.725 \quad \text{(From Table 12)}$$

The two-sided 90% confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{x}_2 - \bar{x}_1) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(85 - 72) \pm 1.725\,(4.48)\sqrt{\frac{1}{10} + \frac{1}{12}}$$

$$13 \pm 3.31 \quad \Rightarrow \quad 9.69 < \mu_2 - \mu_1 < 16.31$$

Exercise 12.4

1. (*a*) A test in statistics was given to 50 girls and 75 boys. The girls made an average grade of 76 with a standard deviation of 6, while the boys made an average grade of 82 with a standard deviation of 8. Find a 96% confidence interval for the difference $\mu_1 - \mu_2$, where μ_1 is the mean score of all boys and μ_2 is the mean score of all girls who might take this test.

$(3.42 < \mu_1 - \mu_2 < 8.58)$

(*b*) A manufacturing company consists of two departments producing identical products. It is suspected that the hourly outputs in the two departments are different. Two random samples of production hours are respectively selected and the following data are obtained:

	Department 1	*Department 2*
Sample size:	64	49
Sample mean:	100	90

The variances of the hourly outputs for the two departments are known to be $\sigma_1^2 = 256$ and $\sigma_2^2 = 196$ respectively. What is the point estimate for the true difference between the mean outputs of the two departments ? Find the 95 percent confidence limits for the true difference.

$(\bar{x}_1 - \bar{x}_2 = 10;\ 4.456 < \mu_1 - \mu_2 < 15.544)$

2. (*a*) Two independent samples of 100 mechanists and 100 carpenters are taken to estimate the difference between the weekly wages of the two categories of workers. The relevant data are given below:

	Sample mean wages	Population variance
Mechanists:	345	196
Carpenters:	340	204

Determine the 95% and the 99% confidence limits for the true difference between the average wages for machinists and carpenters.

$(1.08 < \mu_1 - \mu_2 < 8.92;\ -0.152 < \mu_1 - \mu_2 < 10.152)$

(*b*) General Incorporated Mill's packing machinery is known to pour dry cereal into economy-size boxes with a standard deviation of 0.6 ounce. Two samples taken on two machines yields the following information:

Machine I	*Machine II*
$n_1 = 15$	$n_2 = 21$
$\bar{x}_1 = 18.7$ ounces	$\bar{x}_2 = 21.9$ ounces

Assuming machine I packages a content that is $N(\mu_1, 0.36)$ and machine II packages a content that is $N(\mu_2, 0.36)$, construct a 95% confidence interval estimate of $\mu_2 - \mu_1$.

$(2.8 < \mu_2 - \mu_1 < 3.6)$

3. (*a*) A sample of 150 brand *A* light bulbs showed a mean lifetime of 1400 hours with a standard deviation of 120 hours. A sample of 200 brand *B* light bulbs showed a mean lifetime of 1200 hours with a standard deviation of 80 hours. Find 95% and 99% confidence limits for the difference between the mean lifetime of the populations of brands *A* and *B*.

$(177.825 < \mu_1 - \mu_2 < 222.175;\ 170.856 < \mu_1 - \mu_1 < 229.144)$

(*b*) Let two independent random samples, each of size 100, from independent normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ yield $\bar{x}_1 = 4.8$, $\hat{s}_1^2 = 8.64$, $\bar{x}_2 = 5.6$, $\hat{s}_2^2 = 7$. Find a 95% confidence interval for $(\mu_2 - \mu_1)$.

$(0.025 < \mu_2 - \mu_1 < 1.575)$

4. (*a*) In order to ascertain the age distribution of operatives in a certain industry, random samples of 1720 males and 1230 females are drawn. The sample means and standard deviations were 33.93 years and 14.20 years for the males and 27.44 years and 10.79 years for the females. Calculate the 95 percent confidence interval for

(*i*) the mean age of all male operatives,

(*ii*) the mean age of all female operatives,

(*iii*) the difference between their mean ages.

$(33.259 < \mu_1 < 34.601;\ 26.837 < \mu_2 < 28.043;\ 5.588 < \mu_1 - \mu_2 < 7.392)$

(*b*) The means and variances of the weekly incomes in rupees of the workers employed in the different factories, from the samples are given below:

Sample	*Size*	*Mean*	*Variance*
Factory A	160	12.80	64
Factory B	220	11.25	49

(*i*) What is the maximum likelihood estimate of the difference in mean incomes?

(*ii*) Compute the 95 percent confidence interval estimate for the real differences in the incomes of the workers from the two factories.

{ (*i*) 1.55, (*ii*) $0.003 < \mu_1 - \mu_2 < 3.097$ }

5. (*a*) Let two independent random samples, each of size 100, from two independent normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ yield $\bar{x} = 4.8$, $\hat{s}_1^2 = 8.64$, $\bar{y} = 5.6$, $\hat{s}_2^2 = 7.88$. Find a 95% confidence interval for $(\mu_1 - \mu_2)$.

$(-1.6 < \mu_1 - \mu_2 < 0)$

(*b*) Given that

$$\bar{x}_1 = 75, \quad n_1 = 9, \quad \Sigma(x_{i1} - \bar{x}_1)^2 = 1482$$

$$\bar{x}_2 = 60, \quad n_2 = 16, \quad \Sigma(x_{i2} - \bar{x}_2)^2 = 1830$$

and assuming that the two samples were randomly selected from two normal populations in which $\sigma_1^2 = \sigma_2^2$ (but unknown), calculate an 80% confidence interval for the difference between the two population means.
$(8.4 < \mu_1 - \mu_2 < 21.6)$

6. (*a*) Two random samples of size $n_1 = 9$ and $n_2 = 16$ from two independent population having normal distributions provide the means and standard deviations; $\bar{x}_1 = 64$, $\bar{x}_2 = 59$, $\hat{s}_1 = 6$ and $\hat{s}_2 = 5$. Find a 95% confidence interval for $\mu_1 - \mu_2$ assuming $\sigma_1 = \sigma_2$.
$(0.37 < \mu_1 - \mu_2 < 9.63)$

(*b*) A course in mathematics is taught to 12 students by the conventional class- room method. A second group of 10 students was given the same course by means of programmed materials. At the end of the course, the same examination was given to each group. The 12 students meeting in the class room made an average grade of 85 with a standard deviation of 4, while the 10 students using programmed materials made an average of 81 with a standard deviation of 5. Find a 90% confidence interval for the difference between the population means, assuming the populations to be approximately normally distributed with equal variance.
$(0.693 < \mu_1 - \mu_2 < 7.307)$

◆

12.12 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS, $\pi_1 - \pi_2$

We now turn to statistical inferences concerning a comparison between the rates of incidence of a characteristic into populations. Comparing infant mortality in two groups, the unemployment rates in rural and urban populations, and the proportion of defective items produced by two competing manufacturing processes are the examples of this type. The unknown proportion of elements possessing the particular characteristic in population I and in population II are denoted by π_1 and π_2, respectively. Our aim is to construct confidence intervals for the parameter $\pi_1 - \pi_2$.

A random sample of size n_1 is taken from population I and the number of successes is denoted by X_1. An independent random sample of size n_2 is taken from population II and the number of successes is denoted by X_2. The sample proportions of successes are

$$P_1 = \frac{X_1}{n_1}, \qquad P_2 = \frac{X_2}{n_2}$$

An intuitively appealing estimator for $\pi_1 - \pi_2$ is the difference between the sample proportions $P_1 - P_2$. When constructing the confidence intervals for $\pi_1 - \pi_2$, we will use the sampling distribution of $P_1 - P_2$.

When both sample sizes n_1 and n_2 are large, the Central Limit Theorem assures that P_1 is approximately normal with mean π_1 and variance $\pi_1(1-\pi_1)/n_1$ and that P_2 is approximately normal with mean π_2 and variance $\pi_2(1-\pi_2)/n_2$ and that P_1 is independent of P_2

Since any linear combination of independent normal random variables is also normally distributed then for large sample sizes n_1 and n_2, the sampling distribution of the random variable $P_1 - P_2$ is approximately normal with mean

$$\mu_{P_1 - P_2} = \pi_1 - \pi_2$$

and standard deviation

$$\sigma_{P_1 - P_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

The first result shows that $P_1 - P_2$ is an unbiased estimator of $\pi_1 - \pi_2$. For large sample sizes n_1 and n_2, the random variable

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

is approximately standard normal. The estimate of the standard error of $P_1 - P_2$ can be obtained by replacing π_1 and π_2 by their sample estimates P_1 and P_2 as

$$\hat{\sigma}_{P_1 - P_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

The random variable Z then becomes

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$$

Therefore, in the case of two large, independent random samples a $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ can be readily constructed from this approximation.

Then a two-sided $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is given by

$$(P_1 - P_2) \pm z_{1-\alpha/2}\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

If p_1 and p_2 are the proportions in the two large, independent observed random samples, then a $100(1-\alpha)\%$ interval for $\pi_1 - \pi_2$ is given by

$$(p_1 - p_2) \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

***Example* 12.21** *An antibiotic for pneumonia was injected into* 100 *patients with kidney malfunctions (called uremic patients) and into* 80 *patients with no kidney malfunctions (called normal patients). Some allergic reaction developed in* 40 *of the uremic patients and in* 16 *of the normal patients. Construct a* 95% *confidence interval for the difference between the population proportions.*

Solution. The sizes, number of successes and proportions of successes in the two samples are

$$n_1 = 100, \qquad x_1 = 40, \qquad p_1 = \frac{x_1}{n_1} = \frac{40}{100} = 0.4$$

$$n_2 = 80, \qquad x_2 = 16, \qquad p_2 = \frac{x_2}{n_2} = \frac{16}{80} = 0.2$$

Confidence coefficient: $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{ \text{From Table 10 } (b) \}$$

The two-sided 95% confidence interval for difference in the population proportions $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$(0.4 - 0.2) \pm 1.960 \sqrt{\frac{0.4(1-0.4)}{100} + \frac{0.2(1-0.2)}{80}}$$

$$(0.07, 0.33) \qquad \Rightarrow \qquad 0.07 < \pi_1 - \pi_2 < 0.33$$

Exercise 12.5

1. **(*a*)** In a poll of college students in a large state university, 300 out of 400 students living in dormitories approved a certain course of action, whereas 200 out of 300 students not living in dormitories approved it. Estimate the difference in the proportions favouring the course of action and compute 90% confidence interval for it.
$(p_1 - p_2 = 0.08; \quad 0.023 < \pi_1 - \pi_2 < 0.137)$

(*b*) In a random sample of 400 adults and 600 teenagers who watched a certain television programme. 100 adults and 300 teenagers indicated that they liked it. Construct 95% and 99% confidence limits for the difference in proportions of all adults and all teenagers who watched the programme and liked it.
$(0.19 < \pi_2 - \pi_1 < 0.31; \quad 0.17 < \pi_2 - \pi_1 < 0.33)$

2. **(*a*)** A poll is taken among the residents of a city and the surrounding country to determine the feasibility of a proposal to construct a civic centre. If 2400 of 5000 city residents favour the proposal and 1200 of 2000 country residents favour it, find a 95% confidence interval for the true difference in the proportions favouring the proposal

to construct the civic centre.
$(0.0945 < \pi_2 - \pi_1 < 0.1455)$

(*b*) The population of interest are the voting preferences of all registered voters in the Punjab and the Sind. Two independent random samples were taken from these populations and the values $n_1 = n_2 = 1000$, $p_1 = 0.54$ and $p_2 = 0.47$. Find a 95% confidence interval for $\pi_1 - \pi_2$.
$(0.026 < \pi_1 - \pi_2 < 0.114)$

3. (*a*) A market survey organization carried out a product taste study with consumers in two regions. In one region, a random sample of $n_1 = 400$ consumers was selected while in the other region an independent random sample of $n_2 = 300$ consumers was selected. Each person was asked to indicate which of two servings of product had a better taste. Unknown to the subject, one serving was a new high protein breakfast cereal and the other was an existing cereal. In the first region, proportion $p_1 = 0.55$ of the sample persons preferred the new cereal and in the second region the proportion was $p_2 = 0.65$. Construct a 90% confidence interval for $\pi_2 - \pi_1$.
$(0.039 < \pi_2 - \pi_1 < 0.161)$

(*b*) Independent random samples are selected from two populations with fractions of success π_1 and π_2. Construct a 95% confidence interval for $\pi_1 - \pi_2$ for each of the following cases.

(*i*) $n_1 = 100$ $p_1 = 0.72$ $n_2 = 100$ $p_2 = 0.61$

(*ii*) $n_1 = 130$ $p_1 = 0.16$ $n_2 = 210$ $p_2 = 0.25$

(*iii*) $n_1 = 70$ $p_1 = 0.53$ $n_2 = 60$ $p_2 = 0.48$

{ (*i*) -0.02 to 0.24 (*ii*) -0.176 to -0.004 (*iii*) -0.12 to 0.22 }

◆

Exercise 12.6
Objective Questions

1. Fill in the blanks.

(*i*) Statistical ——— is the conclusion made about the unknown value of population parameter by using the sample observations. (*inference*)

(*ii*) The statistical ——— is a procedure of making judgment about the unknown value of population parameters by using the sample observations. (*estimation*)

(*iii*) The object of ——— estimation is to obtain a single number from the sample that is intended for estimating the unknown true value of a population parameter. (*point*)

(*iv*) A point estimator is a ——— variable whereas an estimate is a constant. (*random*)

(*v*) An estimator is ——— if its expected value is equal to the population parameter to be estimated. (*unbiased*)

(*vi*) If T is a biased estimator, then ———— is the difference of its expected value from the parameter θ to be estimated. (*bias*)

(*vii*) The sample mean $\overline{X}$ is an ———— estimator of population mean μ. (*unbiased*)

(*viii*) The sample proportion P is an ———— estimator of population proportion π. (*unbiased*)

(*ix*) The sample variance $\hat{S}^2 = \Sigma(X - \overline{X})^2/(n-1)$ is an ———— estimator of population variance σ^2. (*unbiased*)

(*x*) ———— estimation is a procedure of constructing an interval from a random sample, such that prior to sampling, it has a high specified probability of including the unknown true value of a population parameter. (*Interval*)

2. Fill in the blanks.

(*i*) The width of a confidence interval is ———— if the level of confidence ($1 - \alpha$) is decreased. (*decreased*)

(*ii*) The width of a confidence interval is ———— related to confidence coefficient. (*directly*)

(*iii*) The precision of confidence interval is increased by ———— the level of confidence. (*decreasing*)

(*iv*) The width of a confidence interval ———— if the sample size is increased. (*decreases*)

(*v*) The confidence coefficient is also called ———— of confidence. (*level*)

(*vi*) $1 - \alpha$ is the ———— that the interval estimator includes the unknown true value of the population parameter. (*probability*)

(*vii*) A sample consisting of 30 or less observations is known as a ———— sample. (*small*)

(*viii*) A sample consisting of more than 30 observations is known as a ———— sample. (*large*)

3. Mark off the following statements as *true* or *false*

(*i*) The types of statistical inferences are estimation of parameters and testing of hypotheses. (*true*)l

(*ii*) The types of statistical estimation of parameters are point estimation and interval estimation. (*true*)

(*iii*) A point estimator is a sample statistic that is used to estimate the unknown true value of a population parameter. (*true*)

(*iv*) A point estimate is a specific value of an estimator computed from the sample data after the sample has been observed. (*true*)

(*v*) An estimate obtained from the sample observations is always a point estimate. (*false*)

(*vi*) Point estimators may be more useful than interval estimators because probability statements are attached to point estimates. (*false*)

(*vii*) The point estimation provides two values with a probability statement for estimating the unknown true value of a population parameter. (*false*)

(*viii*) A confidence interval is a type of statistical inference. (*true*)

(*ix*) A point estimate provides information about the precision of the estimate. (*false*)

4. Mark off the following statements as *true* or *false*

(*i*) We cannot control the precision of an interval estimate by the choice of sample size or level of confidence. (*false*)

(*ii*) The width of a confidence interval increases if the confidence coefficient is decreased. (*false*)

(*iii*) The width of a confidence interval decreases if the confidence coefficient is decreased. (*true*)

(*iv*) The width of a confidence interval can be decreased by decreasing the confidence coefficient. (*true*)

(*v*) The precision of an interval estimate can be increased by decreasing the sample size. (*false*)

(*vi*) The precision of an interval estimate can be increased either by increasing the sample size or by decreasing the confidence coefficient. (*true*)

(*vii*) α is the probability that the interval estimator includes the unknown true value of the population parameter. (*false*)

(*viii*) The statistic T can be used in making confidence interval for μ when population is non-normal. (*false*)

13 HYPOTHESIS TESTING

13.1 THE ELEMENTS OF A TEST OF HYPOTHESIS

We are often concerned with testing or drawing a conclusion about the population parameter θ based on a simple random sample data. In this chapter we present formal structures for making inferences about population parameters such as μ, π, σ^2, *etc.*, we will begin by introducing a *test of hypothesis.*

13.1.1 Statistical Hypothesis. A *statistical hypothesis* is an assertion or conjecture about the distribution of one or more random variables. It is an assertion about the nature of a population. This assertion may or may not be true. Its validity is tested on the basis of an observed random sample. Hypotheses are usually phrased quantitatively in terms of population parameters. The following are some examples of hypotheses.

(*i*) $\mu = 25$ (A population mean equals 25)

(*ii*) $\mu < 40$ (A population mean is less than 40)

(*iii*) $\mu \geq 20$ (A population mean is greater than or equal to 20)

(*iv*) $\pi = 0.4$ (A population proportion equals 0.4)

13.1.2 Hypothesis Testing. The *statistical hypothesis testing* is a procedure to determine whether or not an assumption about some parameter of a population is supported by the information obtained from the observed random sample.

13.1.3 Specification of the Form of Population Distribution. The experimenter, or researcher, must make an assumption about the nature of the underlying distribution of the population. Is the random variable normal or binomial. Or does it follow any other form of the distribution. It is, therefore, necessary to identify the theoretical probability distribution of the random variable under consideration because the decision about the hypothesis is made on the basis of probability of occurrence.

We will find that there are certain elements common to all tests of hypotheses. These elements are introduced and discussed below:

13.1.4 Null Hypothesis. A *null hypothesis*, denoted by H_0, is that hypothesis which is tested for possible rejection (or nullification) under the assumption that it is true.

A null hypothesis is always a statement of either (1) "no effect" or (2) "the status quo", such as the given coin is unbiased, or a drug is ineffective in curing a particular disease, or there is no difference between the two production methods, or the production line requires no preventive maintenance, *etc.* The experiment is conducted to see if this hypothesis is unreasonable.

13.1.5 Establishment of the Null Hypothesis. Let θ represent the true but unknown value of the population parameter and θ_0 a value on the number line, the hypothesis to be tested will take on one of the following three forms.

(*i*) $\theta = \theta_0$, that is, the true value of the population parameter is equal to some specified value θ_0.

(*ii*) $\theta \geq \theta_0$, that is, the true value of the population parameter is equal to or greater than some specified value θ_0.

(*iii*) $\theta \leq \theta_0$, that is, the true value of the population parameter is equal to or less than some specified value θ_0.

13.1.6 Alternative Hypothesis. An *alternative hypothesis,* denoted by H_1, is that hypothesis which we are willing to accept when the null hypothesis is rejected.

An alternative hypothesis gives the opposing conjecture to that given in the null hypothesis. The alternative hypothesis is often called the *research hypothesis,* because this hypothesis expresses the theory that the experimenter, or researcher, believes to be true. The experiment is conducted to see if the alternative hypothesis is supported

13.1.7 Formulation of Null and Alternative Hypothesis. The alternative (or research) hypothesis is a statement about the value of a population parameter that an investigator attempts to support with observed random sample. The statistical hypothesis testing makes use of the null hypothesis that refers to the same population parameter but denies the alternative hypothesis. Thus the basic strategy in statistical hypothesis testing is to attempt to support the research hypothesis by contradicting the null hypothesis. Therefore, when choosing the null and alternative hypotheses, take the following steps:

(*i*) The experiment is conducted to see if there is support for some hypothesis. This will be the alternative hypothesis, expressed as an inequality in the form "less than" or "greeter than" or "not equal to".

Example : H_1: $\theta < 40$

(*ii*) State the null hypothesis with an equality sign as a complement of the alternative hypothesis.

Example : H_0: $\theta \geq 40$

The following table presents the three types of alternative hypotheses that constitute the counterparts to the three types of null hypotheses.

	Null hypothesis H_0	Alternative hypothesis H_1
1.	$\theta \geq \theta_0$	$\theta < \theta_0$
2.	$\theta \leq \theta_0$	$\theta > \theta_0$
3.	$\theta = \theta_0$	$\theta \neq \theta_0$ (*i.e.*, $\theta < \theta_0$ or $\theta > \theta_0$)

***Example* 13.1** *Formulate the null and the alternative hypotheses used in test of hypothesis for each of the following:*

(*i*) *The mean lifetime of electric light bulbs newly manufactured by a company has not changed from the previous mean lifetime of* 1200 *hours.*

(*ii*) *An automobile is driven on the average no more than* 16000 *kilometers per year.*

(*iii*) *At least* 10% *of the people of Pakistan pay income tax.*

(*iv*) *The proportion of the households that do not own a colour television set is more than* 0.40 *in a locality.*

(*v*) *The average yield of corn of variety A exceeds the average yield of variety B by at least* 200 *kilogram per acre.*

Solution. The null hypothesis H_0 and the alternative hypothesis H_1 for each of the given situations are:

(*i*)	$H_0: \mu = 1200$ hours	against	$H_1: \mu \neq 1200$ hours
(*ii*)	$H_0: \mu \leq 16000$ kilometers	against	$H_1: \mu > 16000$ kilometers
(*iii*)	$H_0: \pi \geq 0.10$	against	$H_1: \pi < 0.10$
(*iv*)	$H_0: \pi \leq 0.40$	against	$H_1: \pi > 0.40$
(*v*)	$H_0: \mu_1 - \mu_2 \geq 200$ kg	against	$H_1: \mu_1 - \mu_2 < 200$ kg

13.1.8 Simple Hypothesis. A statistical hypothesis is said to be a *simple hypothesis* if it completely specifies the underlying population distribution, namely

(*i*) The functional form of the distribution, and

(*ii*) The specific values of all of its parameters.

That is, if it specifies the particular member of the particular family of probability distributions, *e. g.*, a random variable has a normal distribution with mean $\mu = 30$ and standard deviation $\sigma = 4$; or a random variable has a binomial distribution with $n = 6$ and $\pi = 0.4$.

13.1.9 Composite Hypothesis. A statistical hypothesis is said to be a *composite hypothesis* if it does not completely specify the underlying population distribution, *e. g.*, a random variable has a normal distribution with mean $\mu = 40$ or a random variable has a normal distribution with mean $\mu = 50$ and standard deviation $\sigma \geq 4$; or random variable has a distribution with mean $\mu = 40$ and standard deviation $\sigma = 5$; or a random variable has a binomial distribution with $\pi = 0.4$.

13.1.10 Test Statistic. The *test statistic* is a sample statistic which provides a basis for deciding whether or not the null hypothesis should be rejected. The most commonly used test statistics are Z, T, χ^2 and F.

13.1.11 Rejection Region. A *rejection region* specifies a set of values of the test statistic for which the null hypothesis is rejected (and for which the alternative hypothesis is accepted). It is also called as the *critical region*.

13.1.12 Nonrejection Region. A *nonrejection region* specifies a set of values of the test statistic for which the null hypothesis is not rejected. It is also called as the *noncritical region*.

13.1.13 Critical Values. The values of the test statistic which separate the rejection and nonrejection regions for the test are called *critical values*.

13.1.5 Establishment of the Null Hypothesis. Let θ represent the true but unknown value of the population parameter and θ_0 a value on the number line, the hypothesis to be tested will take on one of the following three forms.

(*i*) $\theta = \theta_0$, that is, the true value of the population parameter is equal to some specified value θ_0.

(*ii*) $\theta \geq \theta_0$, that is, the true value of the population parameter is equal to or greater than some specified value θ_0.

(*iii*) $\theta \leq \theta_0$, that is, the true value of the population parameter is equal to or less than some specified value θ_0.

13.1.6 Alternative Hypothesis. An *alternative hypothesis*, denoted by H_1, is that hypothesis which we are willing to accept when the null hypothesis is rejected.

An alternative hypothesis gives the opposing conjecture to that given in the null hypothesis. The alternative hypothesis is often called the *research hypothesis*, because this hypothesis expresses the theory that the experimenter, or researcher, believes to be true. The experiment is conducted to see if the alternative hypothesis is supported

13.1.7 Formulation of Null and Alternative Hypothesis. The alternative (or research) hypothesis is a statement about the value of a population parameter that an investigator attempts to support with observed random sample. The statistical hypothesis testing makes use of the null hypothesis that refers to the same population parameter but denies the alternative hypothesis. Thus the basic strategy in statistical hypothesis testing is to attempt to support the research hypothesis by contradicting the null hypothesis. Therefore, when choosing the null and alternative hypotheses, take the following steps:

(*i*) The experiment is conducted to see if there is support for some hypothesis. This will be the alternative hypothesis, expressed as an inequality in the form "less than" or "greeter than" or "not equal to".

Example: H_1: $\theta < 40$

(*ii*) State the null hypothesis with an equality sign as a complement of the alternative hypothesis.

Example: H_0: $\theta \geq 40$

The following table presents the three types of alternative hypotheses that constitute the counterparts to the three types of null hypotheses.

	Null hypothesis H_0	Alternative hypothesis H_1
1.	$\theta \geq \theta_0$	$\theta < \theta_0$
2.	$\theta \leq \theta_0$	$\theta > \theta_0$
3.	$\theta = \theta_0$	$\theta \neq \theta_0$ (*i.e.*, $\theta < \theta_0$ or $\theta > \theta_0$)

***Example* 13.1** *Formulate the null and the alternative hypotheses used in test of hypothesis for each of the following*:

(*i*) *The mean lifetime of electric light bulbs newly manufactured by a company has not changed from the previous mean lifetime of* 1200 *hours.*

(*ii*) *An automobile is driven on the average no more than* 16000 *kilometers per year.*

(*iii*) *At least* 10% *of the people of Pakistan pay income tax.*

(*iv*) *The proportion of the households that do not own a colour television set is more than* 0.40 *in a locality.*

(*v*) *The average yield of corn of variety A exceeds the average yield of variety B by at least* 200 *kilogram per acre.*

Solution. The null hypothesis H_0 and the alternative hypothesis H_1 for each of the given situations are:

(*i*)	$H_0: \mu = 1200$ hours	against	$H_1: \mu \neq 1200$ hours
(*ii*)	$H_0: \mu \leq 16000$ kilometers	against	$H_1: \mu > 16000$ kilometers
(*iii*)	$H_0: \pi \geq 0.10$	against	$H_1: \pi < 0.10$
(*iv*)	$H_0: \pi \leq 0.40$	against	$H_1: \pi > 0.40$
(*v*)	$H_0: \mu_1 - \mu_2 \geq 200$ kg	against	$H_1: \mu_1 - \mu_2 < 200$ kg

13.1.8 Simple Hypothesis. A statistical hypothesis is said to be a *simple hypothesis* if it completely specifies the underlying population distribution, namely

(*i*) The functional form of the distribution, and

(*ii*) The specific values of all of its parameters.

That is, if it specifies the particular member of the particular family of probability distributions, *e. g.*, a random variable has a normal distribution with mean $\mu = 30$ and standard deviation $\sigma = 4$; or a random variable has a binomial distribution with $n = 6$ and $\pi = 0.4$.

13.1.9 Composite Hypothesis. A statistical hypothesis is said to be a *composite hypothesis* if it does not completely specify the underlying population distribution, *e. g.*, a random variable has a normal distribution with mean $\mu = 40$ or a random variable has a normal distribution with mean $\mu = 50$ and standard deviation $\sigma \geq 4$; or random variable has a distribution with mean $\mu = 40$ and standard deviation $\sigma = 5$; or a random variable has a binomial distribution with $\pi = 0.4$.

13.1.10 Test Statistic. The *test statistic* is a sample statistic which provides a basis for deciding whether or not the null hypothesis should be rejected. The most commonly used test statistics are Z, T, χ^2 and F.

13.1.11 Rejection Region. A *rejection region* specifies a set of values of the test statistic for which the null hypothesis is rejected (and for which the alternative hypothesis is accepted). It is also called as the *critical region.*

13.1.12 Nonrejection Region. A *nonrejection region* specifies a set of values of the test statistic for which the null hypothesis is not rejected. It is also called as the *noncritical region.*

13.1.13 Critical Values. The values of the test statistic which separate the rejection and nonrejection regions for the test are called *critical values.*

13.1.14 Two-tailed Test. If the critical region is located equally in both tails of the sampling distribution of test statistic, the test is called a *two-tailed* or *two-sided test.* In such tests, the analyst is concerned with detecting values of the test statistic that are either too large or too small to be consistent with the hypothesis being tested.

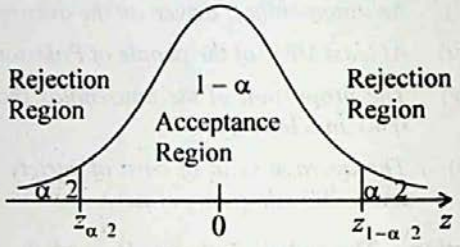


Fig: 13.1 Reject H_0 if $Z < z_{\alpha/2}$ or $Z > z_{1-\alpha/2}$

With a two-tailed test, acceptance of null hypothesis means acceptance of a unique value for the population parameter.

13.1.15 One-tailed Test. If the critical region is located in only one tail of the sampling distribution of test statistic, the test is called a *one-tailed* or *one-sided test.*

13.1.16 Left-tailed Test. If the critical region is located in only the left tail of the sampling distribution of test statistic, the test is called a *left-tailed test.* In such tests, the analyst is concerned with detecting values of the test statistic that are too small to be consistent with the hypothesis being tested.

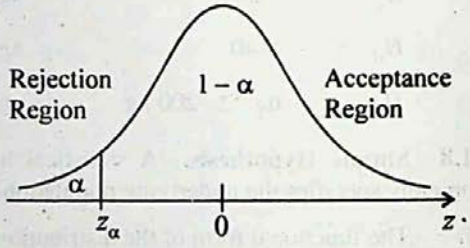


Fig: 13.2 Reject H_0 if $Z < z_\alpha$

With a one-tailed test, acceptance of null hypothesis means accepting that the population parameter is one of many acceptable values.

13.1.17 Right-Tailed Test. If the critical region is located in only the right tail of the sampling distribution of test statistic, the test is called a *right-tailed test.* In such tests the analyst is concerned with detecting the values of the test statistic that are too large to be consistent with the hypothesis being tested.

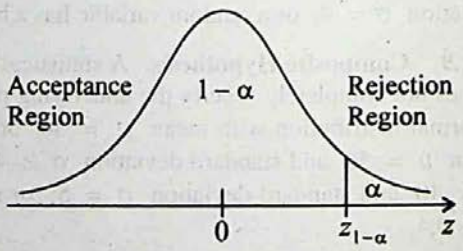


Fig: 13.3 Reject H_0 if $Z > z_{1-\alpha}$

13.1.18 Deciding upon an Appropriate Test. Now the question arises, that how we should decide upon an appropriate test. While deciding upon an appropriate test the following hints are helpful:

(*i*) If we are looking for a definite decrease, *i. e.*, if H_1 is given by $\theta < \theta_0$ we use a one-sided left tail test.

(*ii*) If we are looking for a definite increase, *i. e.*, if H_1 is given by $\theta > \theta_0$ we use a one-sided right tail test.

(*iii*) If we are looking for any change, *i. e.*, if H_1 is given by $\theta \neq \theta_0$ we use a two-sided test.

The preceding discussion describing the nature of the one-tailed and two-tailed tests is summarized in the following table:

Table : Describing the nature of test.

Null hypothesis	Alternative hypothesis	Type of test	Location of rejection region
$\theta \geq \theta_0$	$\theta < \theta_0$	Left-tailed	The rejection region is located in the left tail of the sampling distribution under H_0.
$\theta \leq \theta_0$	$\theta > \theta_0$	Right-tailed	The critical region is located in the right tail of the sampling distribution under H_0.
$\theta = \theta_0$	$\theta \neq \theta_0$	Two-tailed	The rejection region is located equally in both tails of the sampling distribution under H_0.

Remarks : The names of the three types of a test are associated with alternative hypothesis H_1.

The choice of a one-sided left tail test, one-sided right tail test or a two-sided test depends upon the type of alternative hypothesis.

13.1.19 Errors of Inference. Sample evidence is used to test the null hypothesis H_0. If the sample evidence convinces us that H_0 has only a small chance of being true (a large chance of being false), we say H_0 is unreasonable and reject it. If the sample evidence does not convince us that H_0 is unreasonable, we do not reject H_0. Therefore, there are two alternative conclusions.

(*i*) Reject the null hypothesis on the basis of sample evidence.

(*ii*) Accept the null hypothesis on the basis of sample evidence.

There are two possible states of nature of the null hypothesis.

(*i*) The null hypothesis is true.

(*ii*) The null hypothesis is false.

Thus, in hypothesis testing one and only one of the following four possible outcomes will occur.

(*i*) If the null hypothesis is true, we may reject it leading to a wrong decision.

(*ii*) If the null hypothesis is true, we may accept or not reject it leading to a correct decision.

(*iii*) If the null hypothesis is false, we may accept it leading to a wrong decision.

(*iv*) If the null hypothesis is false, we may reject it leading to a correct decision.

Type-I Error. A Type-I error is made by rejecting H_0 if H_0 is actually true.

Type-II Error. A Type-II error is made by accepting H_0 if H_1 is actually true.

These four possible outcomes can be displayed in four cells of a table as follows:

Table : Hypothesis and conclusion reached from sample

	State of nature of hypotheses	
Sample indicates conclusion	H_0 is true	H_1 is true
Reject H_0	Type-I error	Correct decision
Do not reject (or accept) H_0	Correct decision	Type-II error

The probabilities of the four possible outcomes are commonly designated as

$$\alpha = P(\text{Rejecting } H_0 \,|\, H_0 \text{ is true}) = P(\text{Type-I error})$$

$$1 - \alpha = P(\text{Accepting } H_0 \,|\, H_0 \text{ is true})$$

$$\beta = P(\text{Accepting } H_0 \,|\, H_1 \text{ is true}) = P(\text{Type-II error})$$

$$1 - \beta = P(\text{Rejecting } H_0 \,|\, H_1 \text{ is true})$$

We can represent these conditional probabilities as:

Table : Conditional probabilities

	State of nature of hypotheses	
Sample indicates conclusion	H_0 is true	H_1 is true
Reject H_0	α	$1 - \beta$
Do not reject (or accept) H_0	$1 - \alpha$	β

13.1.20 Level of Significance. The *level of significance* of a test is the maximum probability with which we are willing to a risk of Type-I error. It is the probability of obtaining a value of the test statistic inside the critical region given that H_0 is true, *i. e.*, it is the probability of rejecting a true null hypothesis. It is denoted by α.

$$P(\text{Rejecting } H_0 \,|\, H_0 \text{ is true}) = P(\text{Type-I error}) = \alpha$$

The level of significance is also called as the *size of the critical region* or the *size of the test.* It is a small pre-assigned value, say, 0.05 or 0.01, which is generally specified before any samples are drawn, so that the results obtained will not influence our choice.

13.1.21 Level of Confidence. The *level of confidence* is the probability of accepting a true null hypothesis. It is denoted by $1 - \alpha$.

$$P(\text{Accepting } H_0 \,|\, H_0 \text{ is true}) = 1 - \alpha$$

***Example* 13.2** *The prosecuting attorney in a trial attempts to show that the defendant is guilty. The trial can be thought of a test of hypothesis, since a decision is to be made as to whether the defendant is guilty or innocent.*

(*a*) *State the null and alternative hypothesis of interest to the prosecuting attorney.*

(*b*) *Define the Type-*I *error and Type-*II *error for this situation.*

Solution. Since the prosecuting attorney wants to show that the defendant is guilty, this specifies the alternative hypothesis and the hypotheses are

H_0: Defendant is innocent H_1: Defendant is guilty

Type-I error would occur if the defendant were found guilty if, infect, the defendant is innocent.

Type-II error would occur if the defendant were found innocent if, infect, the defendant is guilty.

13.1.22 Statistical Decision Rule. A *statistical decision rule specifies,* for each possible sample outcome, whether the null hypothesis should be rejected or not. It is also called as a *decision function* or a *decision criterion.*

13.1.23 Conclusion. A test of hypothesis leads to one of two conclusions.

(*i*) If the observed value of the test statistic falls in the rejection region, the null hypothesis H_0 is rejected in favour of alternative hypothesis H_1.

(*ii*) If the observed value of the test statistic does not fall in the rejection region the null hypothesis is neither accepted nor rejected. It is, then, stated that there is insufficient evidence to make a decision.

13.1.24 Test of significance. The *test of significance* describes a process of testing a hypothesis to gain a general impression about the parameter. No decision is imminent or even implied. In this case, a level of significance is not present. Instead we investigate the values of α that would lead to rejection of H_0 as opposed to the α values leading to acceptance of H_0. The test of significance differs with hypothesis testing that refers to the act of actually testing a hypothesis at a selected level of significance to make a decision based on that conclusion.

13.1.25 Steps to Follow when Testing a Hypothesis. Since we will be conducting many tests of hypothesis, it is useful to follow a set procedure. In the remainder of this text, we will always follow the same format. The steps we will follow are given below:

***Before any sample observations are considered*:**

1. Identify the population of interest and state the conditions required for the validity of the test procedure being used.
2. Formulate and state the null hypothesis H_0 and the alternative hypothesis H_1.
3. Decide and specify the level of significance, α.
4. Select the appropriate test statistic and its sampling distribution if H_0 is assumed to be true.
5. Give the critical value (or values) of test statistic for desired value of α.
6. Establish the rejection (critical) region.
7. State the decision rule: Reject H_0 if the value of the test statistic from the observed sample falls in the rejection region, otherwise do not reject (or accept) H_0.

***Now consider the sample values*:**

8. Calculate the value of the test statistic from the observe sample.
9. In the light of your decision rule, draw the conclusion as to whether to reject or accept H_0 and then state the decision in managerial terms.

Exercise 13.1

1. (*a*) Define the following concepts in your own words as fully as you can:

(*i*) Hypothesis testing (*ii*) Statistical hypothesis

(*iii*) Null hypothesis (*iv*) Test statistic

(*v*) Level of significance (*vi*) Critical region

(*b*) Explain with examples the difference between:

(*i*) Estimation and hypothesis testing

(*ii*) Null hypothesis and alternative hypothesis

(*iii*) Simple hypothesis and composite hypothesis

(*iv*) Acceptance region and rejection region

(*v*) Type-I error and Type-II error

(*vi*) One-tailed test and two-tailed test

(*c*) What is the difference between a one-sided and a two-sided test? When should each be used?

2. (*a*) Explain what is meant by:

(*i*) Statistical hypothesis (*ii*) Test-statistic

(*iii*) Significance level (*iv*) Test of significance

(*b*) Distinguish between the following concepts:

(*i*) Statistical estimation and hypothesis testing

(*ii*) Rejection and non-rejection regions

(iii) A test at α level of significance and at $1 - \alpha$ confidence level

3. (*a*) Explain how the null hypothesis and the alternative hypothesis are formulated.

(*b*) State the null and alternative hypotheses to be used in testing the following claims:

(*i*) The mean rainfall at Lake Placid during the month of June is 2.8 inches.

(*ii*) No more than 20% of the faculty at Highland University contributed to the annual giving fund.

(*iii*) The proportion of voters favouring Senator Foghorn in up coming election is 0.58.

(*iv*) On the average children attend school within 3.8 miles of their homes in suburban San Francisco.

{ (*i*) H_0: $\mu = 2.8$ inches against H_1: $\mu \neq 2.8$ inches; (*ii*) H_0: $\pi \leq 0.20$ against H_1: $\pi > 0.20$; (*iii*) H_0: $\pi = 0.58$ against H_1: $\pi \neq 0.58$; (*iv*) H_0: $\mu \leq 3.8$ miles against H_1: $\mu > 3.8$ miles.}

4. (*a*) Indicate Type-I and Type-II errors in the following statements:

(*i*) An innocent driver may be held by a traffic constable.

(*ii*) A judge can acquit a guilty person.

(*iii*) A deserving player may not be selected in the team.

(*iv*) A bad student may be passed by the examiner.

{ (*i*) Type-I, (*ii*) Type-II, (*iii*) Type-I, (*iv*) Type-II.}

(*b*) For each of the following situations, indicate the Type-I and Type-II errors and the correct decisions.

(*i*) H_0: New system is no better than the old one.

Adopt new system when new one is better.

Retain old system when new one is better.

Retain old system when new one is not better.

Adopt new system when new one is not better

(*ii*) H_0: New product is satisfactory.

Market new product when unsatisfactory.

Do not market new product when unsatisfactory.

Do not market new product when satisfactory.

Market new product when satisfactory.

{ (*i*) Correct, Type-II, Correct, Type-I; (*ii*) Type-II, Correct, Type-I, Correct }

(*c*) Suppose that a psychological testing service is asked to check whether an executive is emotionally fit to assume the presidency of a large company. What type of error is committed if the hypothesis that he is fit for the job is erroneously accepted? What type of error is committed if the hypothesis that he is fit for the job is erroneously rejected ? (Type-II error; Type-I error)

5. (*a*) The director of an advertising agency is concerned with the effectiveness of a certain kind of television commercial.

(*i*) What hypothesis is he testing, if he is committing a Type-I error when he says erroneously that the commercial is effective.

(*ii*) What hypothesis is he testing, if he is committing a Type-II error when he says erroneously that the commercial is effective.

{ (*i*) Commercial is not effective; (*ii*) Commercial is effective }

(*b*) Outline the fundamental procedure followed in testing a null hypothesis.

◆

13.2 TEST OF HYPOTHESIS ABOUT A POPULATION MEAN, μ

13.2.1 Forms of Hypothesis. Let μ_0 be the hypothesized value of a population mean. The three possible null hypotheses about a population mean, and their corresponding alternative hypotheses, are:

1. $H_0: \mu \geq \mu_0$ against $H_1: \mu < \mu_0$
2. $H_0: \mu \leq \mu_0$ against $H_1: \mu > \mu_0$
3. $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$

In each case, the test will be made by obtaining a simple random sample of size n and computing the sample mean $\overline{X}$. Then $\overline{X}$ will be used in computing a test statistic. Depending upon the calculated value of the test statistic, H_0 will be accepted or rejected.

13.2.2 Normal Population, σ^2 known. We know that for a normal population having a mean μ_0 and a known variance σ^2 the sampling distribution of $\overline{X}$ is normal with mean μ_0 and a standard error of $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Then the statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

has the standard normal distribution. Consequently, the statistic Z is the test statistic for testing a hypothesis about the mean of a normal population whose variance σ^2 is known.

13.2.3 Normal Population, σ^2 unknown. If the variance of the population σ^2 is unknown, then we estimate σ^2 by the sample variance $\hat{S}^2$ and we estimate the standard error of $\overline{X}$ by $S_{\overline{X}}$, where $S_{\overline{X}} = \hat{\sigma}_{\overline{X}} = \hat{S}/\sqrt{n}$. If the population is normal, the statistic

$$T = \frac{\overline{X} - \mu_0}{S_{\overline{X}}} = \frac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}}$$

has a student's t-distribution with $\nu = n - 1$ degrees of freedom. Consequently the statistic T is the test statistic for testing a hypotheses about the mean of a normal population whose variance σ^2 is unknown.

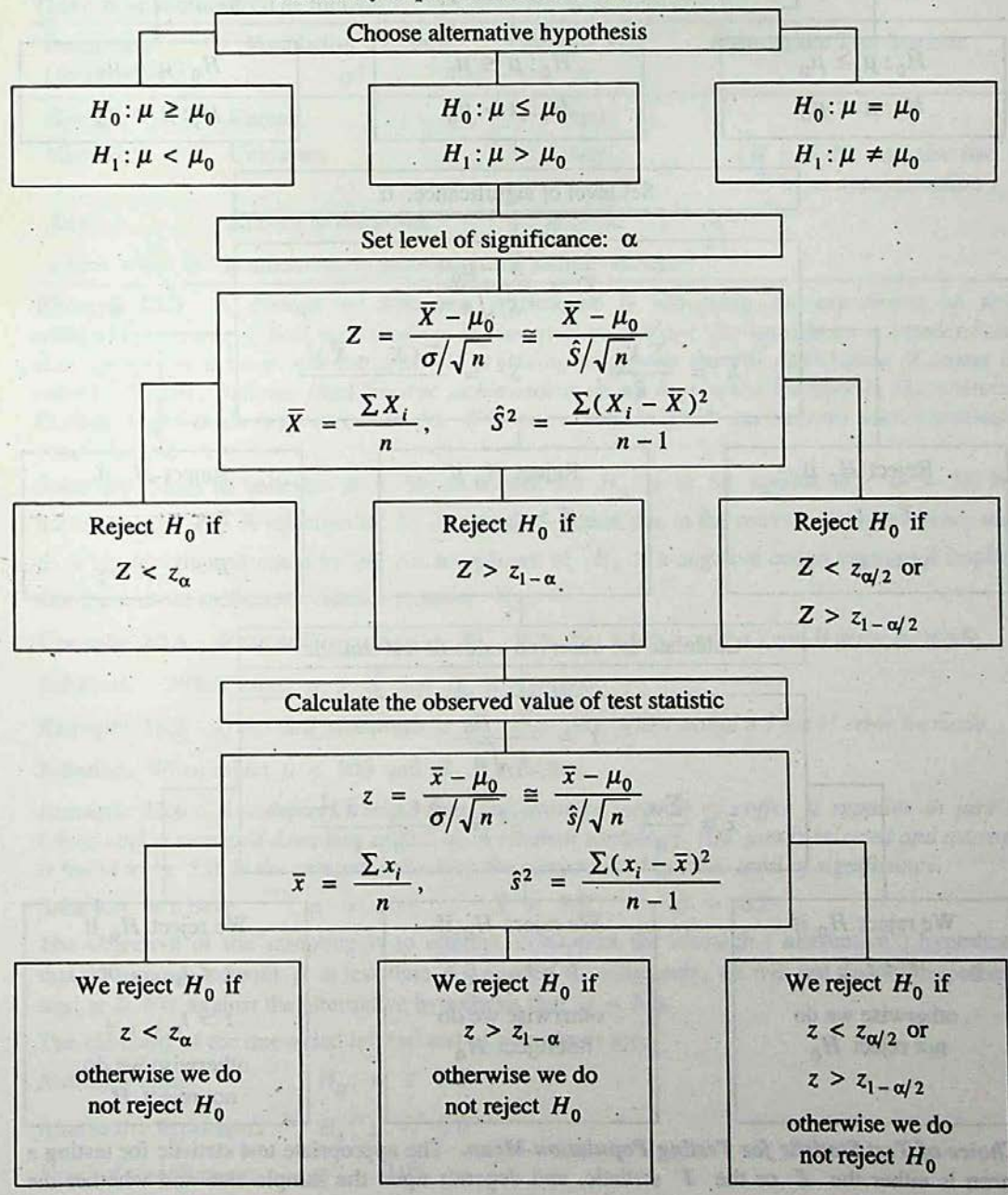
13.2.4 Any Population, σ^2 known/unknown. Finally recall the Central Limit Theorem which states that the sampling distribution of $\overline{X}$ is approximately normal even for non-normal populations if the sample size is sufficiently large (say, $n > 30$). Then we will use the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \cong \frac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}}$$

to test a hypothesis about the mean of any population (normal or not).

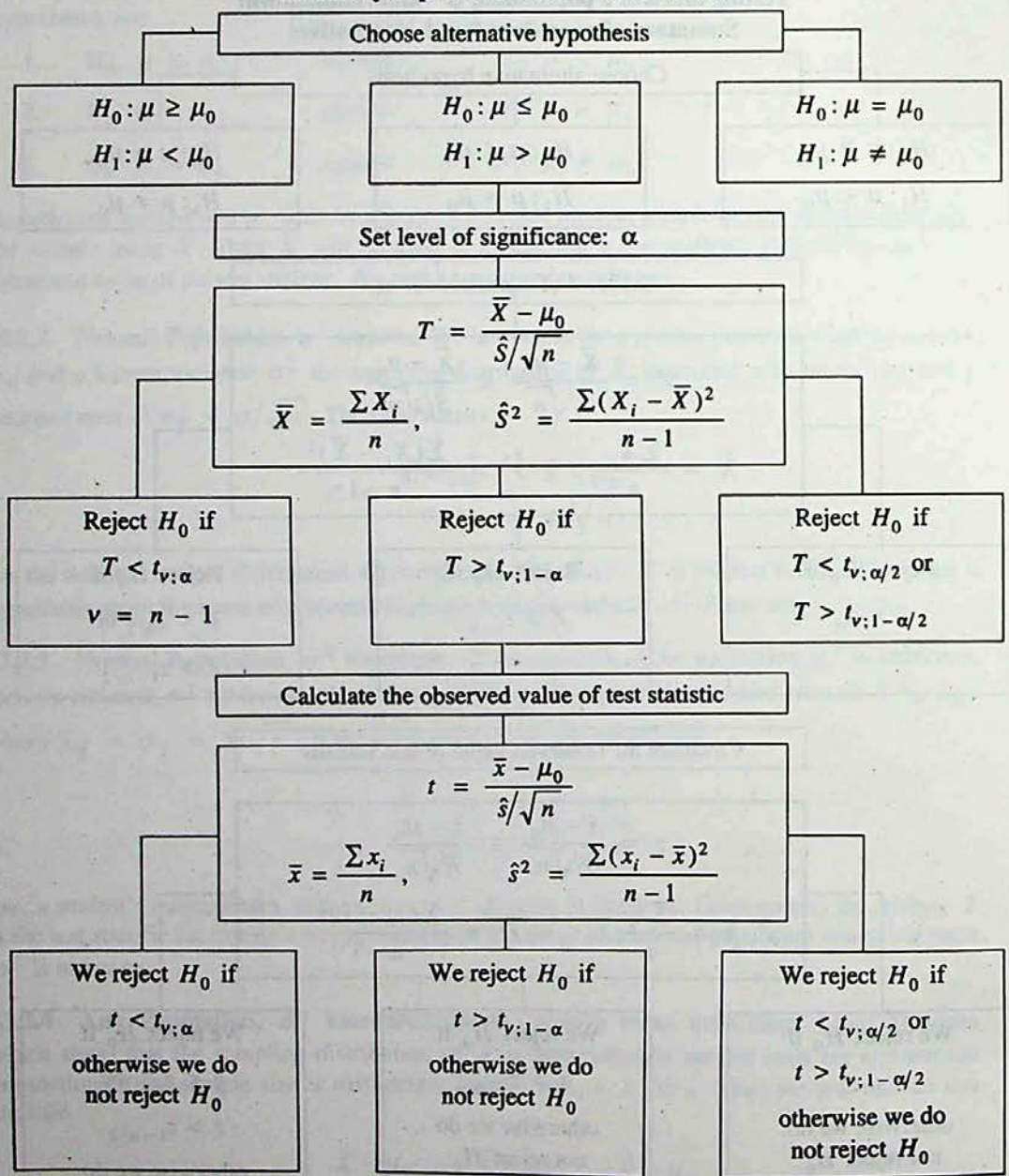
The summary of the procedure of testing a hypothesis about the mean of a population whose variance σ^2 is known/unknown, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing mean of a population, σ^2 known/unknown:
Summary of procedure for 3 alternatives



The summary of the procedure of testing a hypothesis about the mean of a normal population whose variance σ^2 is unknown, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing mean of a normal population, σ^2 unknown: Summary of procedure for 3 alternatives



Choice of Test Statistic for Testing Population Mean. The appropriate test statistic for testing a mean is either the Z or the T statistic, and depends upon the sample size and whether the

population distribution and population variance are known. Thus in testing the hypothesis about the population mean we will use:

(*i*) The test statistic Z when the population is normal and variance is known, or the sample size is sufficiently large,

(*ii*) The test statistic T when the population is normal, whose variance is unknown and sample size is small ($n \leq 30$).

This can be summarized as follows.

Population Distribution	Population Variance σ^2	Sample Size n	Appropriate Test Statistic
Normal	Known	Any	Z
Normal	Unknown	Any	T (If $n > 30$, can also use Z as an approximation)
Any	Known or unknown	$n > 30$	Z

where, when σ^2 is unknown, its estimate is the sample variance $\hat{S}^2$.

***Example* 13.3** *A Bureau of Research statistician is designing an experiment to test achievement scores of first year students in Government College Gujranwala on a standardized test, scored one through one hundred. He is willing to assume that the distribution of scores is normal. He also believes that the true achievement of all first years students in Government College, Gujranwala is greater than* 58. *If he has a sample of* 36 *such scores what hypothesis should he test.*

Solution. Since he believes $\mu > 58$, he should test $H_0: \mu \leq 58$ against $H_1: \mu > 58$. He hopes to reject H_0. A rejection of H_0 is a positive action due to the overwhelming evidence that μ is not less than or equal to 58. An acceptance of H_0 is a negative action because it implies that there is not sufficient evidence to reject H_0.

***Example* 13.4** *If the null hypothesis is* $H_0: \mu \leq 50$, *when would a Type-II error be made.*

Solution. When infact $\mu > 50$ and H_0 is accepted.

***Example* 13.5** *If the null hypothesis is* $H_0: \mu \geq 100$, *when would a Type-II error be made.*

Solution. When infact $\mu < 100$ and H_0 is accepted.

***Example* 13.6** *A company claims that the average amount of coffee it supplies in jars is* 6.0 *oz with a standard deviation of* 0.2 *oz. A random sample of* 100 *jars is selected and average is found to be* 5.9. *Is the company cheating the customers? Use* 5% *level of significance.*

Solution. We have $n = 100, \quad \bar{x} = 5.9, \quad \sigma = 0.2$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the average amount μ is less than 6.0 ounces. Consequently, we will test the null hypothesis that $\mu \geq 6.0$ against the alternative hypothesis that $\mu < 6.0$.

The elements of the one-sided left tail test of hypothesis are:

Null hypothesis $H_0: \mu \geq 6.0$

Alternative hypothesis $H_1: \mu < 6.0$

Level of significance: $\alpha = 0.05$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ follows a standard normal distribution under H_0

Critical values: $z_\alpha = z_{0.05} = -1.645$ { From Table 10 (*b*) }

Critical region: $Z < -1.645$

Decision rule: Reject H_0 if $Z < -1.645$, otherwise do not reject H_0

Observed value: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{5.9 - 6.0}{0.2/\sqrt{100}} = -5.0$

Conclusion: Since $z = -5 < -1.645$, we reject H_0 and conclude that the average amount of coffee is less than 6.0 ounces.

***Example* 13.7** *A company makes parachutes. The company has been buying snap links from a manufacturing firm. The company is concerned that the quality of snap links they receive from the firm might not be up to specifications. Specifically, the company wants to be convinced that snap links will withstand a mean breaking strength of more than* 5000 *pounds. Perform a test of hypothesis at the* 0.005 *significance level if the mean breaking strength for a random sample of* 50 *snap links is* 5100 *pounds. The population standard deviation is* 221 *pounds. What is implied by the test result.*

Solution. We have $n = 50, \quad \bar{x} = 5100, \quad \sigma = 221$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean breaking strength μ is more than 5000 pounds. Consequently, we will test the null hypothesis that $\mu \leq 5000$ against the alternative hypothesis that $\mu > 5000$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu \leq 5000$

Alternative hypothesis $H_1: \mu > 5000$

Level of significance: $\alpha = 0.005 \Rightarrow 1 - \alpha = 1 - 0.005 = 0.995$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ follows a standard normal distribution under H_0.

Critical value: $z_{1-\alpha} = z_{0.995} = 2.576$ { From Table 10 (*b*) }

Critical region: $Z > 2.576$

Decision rule: Reject H_0 if $Z > 2.576$, otherwise do not reject H_0.

Observed value: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{5100 - 5000}{221/\sqrt{50}} = 3.20$

Conclusion: Since $z = 3.20 > 2.576$, we reject H_0 and conclude that the mean breaking strength is more than 5000 pounds. Links from the firm are up to specifications. Buy them.

***Example* 13.8** *The mean lifetime of electric light bulbs produced by a company has in the past been* 1120 *hours. A random sample of* 36 *electric bulbs recently chosen from a supply of newly manufactured bulbs showed a mean lifetime of* 1087 *hours with a standard deviation of* 120

hours. Test the hypothesis that the mean lifetime of light bulbs has not changed using a level of significance of 0.05.

Solution. We have $n = 36, \quad \bar{x} = 1087, \quad \hat{s} = 120$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean lifetime μ has changed from 1120 hours. Consequently, we will test the null hypothesis that $\mu = 1120$ against the alternative hypothesis that $\mu \neq 1120$.

The elements of the two-sided test of hypothesis are:

Null hypothesis $H_0: \mu = 1120$

Alternative hypothesis $H_1: \mu \neq 1120$

Level of significance: $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}}$ follows an approximate standard normal distribution under H_0.

Critical values: $z_{\alpha/2} = z_{0.025} = -1.960,$

$z_{1-\alpha/2} = z_{0.975} = 1.960$ { From Table 10 (*b*) }

Critical region: $Z < -1.960$ or $Z > 1.960$

Decision rule: Reject H_0 if $Z < -1.960$ or $Z > 1.960$, otherwise do not reject H_0.

Observed value: $z = \dfrac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \dfrac{1087 - 1120}{120/\sqrt{36}} = -1.65$

Conclusion: Since $-1.960 < z = -1.65 < 1.960$, we do not reject H_0 and conclude that the mean lifetime is 1120 hours.

***Example* 13.9** *We wish to test the hypothesis that the mean weight of a population of people is* 140 *lb. Using* $\sigma = 15$ *lb.* $\alpha = 0.05$ *and a sample of* 36 *people, find. the values of* $\bar{X}$ *which would lead to rejection of the hypothesis.*

Solution. We have $n = 36, \quad \sigma = 15.$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean weight μ is not equal to 140 pounds. Consequently, we will test the null hypothesis that $\mu = 140$ against the alternative hypothesis that $\mu \neq 140$.

The elements of the two-sided test of hypothesis are:

Null hypothesis $H_0: \mu = 140$

Alternative hypothesis $H_1: \mu \neq 140$

Level of significance: $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ follows a standard normal distribution under H_0

Critical values: $z_{\alpha/2} = z_{0.025} = -1.960,$

$z_{1-\alpha/2} = z_{0.975} = 1.960$ { From Table 10 (*b*) }

Critical region: $Z < -1.960$ or $Z > 1.960$

Decision rule: Reject H_0 if $Z < -1.960$ or $Z > 1.960$

$$\frac{\bar{X} - 140}{15/\sqrt{36}} < -1.960 \quad \text{or} \quad \frac{\bar{X} - 140}{15/\sqrt{36}} > 1.960$$

$$\bar{X} < 135.1 \quad \text{or} \quad \bar{X} > 144.9$$

Hence, the hypothesis H_0: $\mu = 140$ will be rejected if either $\bar{X} < 135.1$ or $\bar{X} > 144.9$.

***Example* 13.10** *An ultrasonic equipment manufacturer uses a component in one of its machines that must withstand considerable stress from vibrations while the machine is operating. An all-metal component has been available for some years from a supplier. Extensive historical experience has shown this component has a mean service life of 1100 hours. The research division of the supplier has just developed a modified component constructed from plastic and metal, bonded in a special way. The manufacturer wishes to know whether or not the mean service life of the modified component (μ) exceeds the mean service life of 1100 hours for the original component. Test the hypothesis at 1% level of significance if a sample of $n = 36$ components yielded $\bar{x} = 1121$ and $\hat{s} = 222$ hours.*

Solution. We have $n = 36$, $\bar{x} = 1121$, $\hat{s} = 222$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean service life μ exceeds 1100 hours. Consequently, we will test the null hypothesis that $\mu \leq 1100$ against the alternative hypothesis that $\mu > 1100$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis H_0: $\mu \leq 1100$

Alternative hypothesis H_1: $\mu > 1100$

Level of significance: $\alpha = 0.01 \Rightarrow 1 - \alpha = 1 - 0.01 = 0.99$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}}$ follows a approximate standard normal distribution under H_0.

Critical value: $z_{1-\alpha} = z_{0.99} = 2.326$ { From Table 10 (*b*) }

Critical region: $Z > 2.326$

Decision rule: Reject H_0 if $Z > 2.326$, otherwise do not reject H_0.

Observed value: $z = \dfrac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \dfrac{1121 - 1100}{222/\sqrt{36}} = 0.568$

Conclusion: Since $z = 0.568 < 2.326$, so we do not reject H_0 and conclude that the mean service life does not exceeds 1100 hours.

***Example* 13.11** *A lumber company is interested in seeing if the number of board feet per tree has decreased since moving to a new location of timber. In the past, the company has an average of 93 board feet per tree. The company believes that the production has decreased since changing locations, a random sample of 25 trees yields $\bar{x} = 89$ with $\hat{s} = 20$. Assuming the normality of the data, test the hypothesis at a 10% level of significance.*

Solution. We have $n = 25, \quad \bar{x} = 89, \quad \hat{s} = 20$.

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean production μ has decreased from 93 board feet. Consequently, we will test the null hypothesis that $\mu \geq 93$ against the alternative hypothesis that $\mu < 93$.

The elements of the one-sided left tail test of hypothesis are:

Null hypothesis $H_0: \mu \geq 93$

Alternative hypothesis $H_1: \mu < 93$

Level of significance: $\alpha = 0.10$

Test statistic: $T = \dfrac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 25 - 1 = 24$

Critical value: $t_{\nu;\alpha} = t_{24;0.10} = -1.318$ (From Table 12)

Critical region: $T < -1.318$

Decision rule: Reject H_0 if $T < -1.318$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \dfrac{89 - 93}{20/\sqrt{25}} = -1.000$

Conclusion: Since $t = -1.000 > -1.318$, we do not reject H_0 and conclude that the mean production has not decreased from 93 board feet. The tree company's claim has not been established.

***Example* 13.12** *Expensive test borings were made in an oil shale area to determine if the mean yield of oil per ton of shale rock is greater than 4.5 barrels. Five borings, made at randomly selected points in the area, indicated the following number of barrels per ton 4.8, 5.4, 3.9, 4.9, 5.5. Suppose barrels per ton are normally distributed. Perform a test of hypothesis at the 5% level of significance.*

Solution. The mean and standard deviation of the sample are

x_i	4.8	5.4	3.9	4.9	5.5	$\sum x_i = 24.5$
x_i^2	23.04	29.16	15.21	24.01	30.25	$\sum x_i^2 = 121.67$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{24.5}{5} = 4.9$$

$$\hat{s} = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{121.67 - 5(4.9)^2}{5-1}} = 0.64$$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean yield μ is greater than 4.5 barrels. Consequently, we will test the null hypothesis that $\mu \leq 4.5$ against the alternative hypothesis that $\mu > 4.5$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu \leq 4.5$

Alternative hypothesis $H_1: \mu > 4.5$

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $T = \dfrac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 5 - 1 = 4$

Critical value: $t_{\nu;\,1-\alpha} = t_{4;\,0.95} = 2.132$ (From Table 12)

Critical region: $T > 2.132$

Decision rule: Reject H_0 if $T > 2.132$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\overline{x} - \mu_0}{\hat{s}/\sqrt{n}} = \dfrac{4.9 - 4.5}{0.64/\sqrt{5}} = 1.40$

Conclusion: Since $t = 1.40 < 2.132$, we do not reject H_0 and conclude that the mean yield is not greater than 4.5 barrels..

***Example* 13.13** *A cattle rancher has changed the type of feed he uses to fatten his cattle for sale. The feed company claims that the new feed will increase the mean weight gain in his cattle by more than* 100 *pounds per steer. Assuming the weight gain of cattle is normally distributed, test the hypothesis of the feed company at* $\alpha = 0.05$. *Previously, the mean weight gain per steer has been* 800 *pounds. A random sample of* 30 *yields a mean weight gain of* $\overline{x} = 935$ *pounds with a standard deviation of* 85 *pounds.*

Solution. We have $n = 30, \quad \overline{x} = 935, \quad \hat{s} = 85$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean weight μ is more than 900 (*i. e.*, 800 + 100) pounds. Consequently, we will test the null hypothesis that $\mu \leq 900$ against the alternative hypothesis that $\mu > 900$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu \leq 900$

Alternative hypothesis $H_1: \mu > 900$

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $T = \dfrac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 30 - 1 = 29$

Critical value: $t_{\nu;\,1-\alpha} = t_{29;\,0.95} = 1.699$ (From Table 12)

Critical region: $T > 1.699$

Decision rule: Reject H_0 if $T > 1.699$, otherwise do not reject H_0.

Observed value: $$t = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{935 - 900}{85/\sqrt{30}} = 2.26$$

Conclusion: Since $t = 2.26 > 1.699$, so we reject H_0 and conclude that the mean weight is more than 900 (*i. e.*, 800 + 100) pounds.. The feed company's claim has been established.

***Example* 13.14** *Workers at a production facility are required to assemble a certain part in 2.3 minutes in order to meet production criteria. The assembly rate per part is assumed to be normally distributed. Six workers are selected at random and timed in assembling a part. The assembly times (in minutes) for the six workers as follows.*

2.0 2.4 1.7 1.9 2.8 1.8

The manager wants to determine if the mean for all workers differs from 2.3. Perform a test of hypothesis at the 5% level of significance.

Solution. The mean and standard deviation of the sample are

x_i	2.0	2.4	1.7	1.9	2.8	1.8	$\Sigma x_i = 12.6$
$x_i - \bar{x}$	−0.1	0.3	−0.4	−0.2	0.7	−0.3	
$(x_i - \bar{x})^2$	0.01	0.09	0.16	0.04	0.49	0.09	$\Sigma(x_i - \bar{x})^2 = 0.88$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{12.6}{6} = 2.1$$

$$\hat{s} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{0.88}{6 - 1}} = 0.42$$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean assembly time μ differs from 2.3 minutes. Consequently, we will test the null hypothesis that $\mu = 2.3$ against the alternative hypothesis that $\mu \neq 2.3$.

The elements of the two-sided test of hypothesis are:

Null hypothesis $H_0: \mu = 2.3$

Alternative hypothesis $H_1: \mu \neq 2.3$

Level of significance: $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$

Test statistic: $$T = \frac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}}$$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 6 - 1 = 5$

Critical values: $t_{\nu;\alpha/2} = t_{5;0.025} = -2.571$,

$t_{\nu;1-\alpha/2} = t_{5;0.975} = 2.571$ (From Table 12)

Critical region: $T < -2.571$ or $T > 2.571$

Decision rule: Reject H_0 if $T < -2.571$ or $T > 2.571$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \dfrac{2.1 - 2.3}{0.42/\sqrt{6}} = -1.166$

Conclusion: Since $-2.571 < t = -1.166 < 2.571$, we do not reject H_0 and conclude that the average assembly time is 2.3 minutes.

Exercise 13.2

1. **(*a*)** Why is the z-test usually inappropriate as a test-statistic when the sample size is small?

(*b*) Define "Student's t-statistic". What are its assumptions? Explain briefly its use and importance in statistics.

2. For each of the following, a random sample of size n is taken from a normal distribution with mean μ and variance σ^2. The sample mean is $\bar{x}$. Test the hypothesis stated, at the level of significance indicated.

Sample	n	$\bar{x}$	σ	Hypotheses		Level of significance
				H_0	H_1	
(*i*)	30	15.2	3	$\mu = 15.8$	$\mu \neq 15.8$	5%
(*ii*)	10	27	1.2	$\mu \leq 26.3$	$\mu > 26.3$	5%
(*iii*)	49	125	4.2	$\mu \leq 123.5$	$\mu > 123.5$	1%
(*iv*)	100	4.35	0.18	$\mu \geq 4.40$	$\mu < 4.40$	2%

Sample	n	$\bar{x}$	$\sum(x_i - \bar{x})^2$	Hypotheses		Level of significance
				H_0	H_1	
(*v*)	65	100	842.4	$\mu = 99.2$	$\mu \neq 99.2$	5%
(*vi*)	65	100	842.4	$\mu \leq 99.2$	$\mu > 99.2$	5%
(*vii*)	80	85.3	2508.8	$\mu \geq 86.2$	$\mu < 86.2$	10%
(*viii*)	100	6.85	36	$\mu = 7.0$	$\mu \neq 7.0$	1%

{ (*i*) Since $z_{0.025} = -1.96 < z = -1.095 < 1.96 = z_{0.975}$, we do not reject $H_0: \mu = 15.8$ against $H_1: \mu \neq 15.8$.

(*ii*) Since $z = 1.845 > 1.645 = z_{0.95}$, we reject $H_0: \mu \leq 26.3$ in favour of $H_1: \mu > 26.3$.

(*iii*) Since $z = 2.5 > 2.326 = z_{0.99}$, we reject $H_0: \mu \leq 123.5$ in favour of $H_1: \mu > 123.5$.

(*iv*) Since $z = -2.778 < -2.054$, we reject $H_0: \mu \geq 4.40$ in favour of $H_1: \mu < 4.40$.

(*v*) Since $z_{0.025} = -1.96 < z = 1.778 < 1.96 = z_{0.975}$, we do not reject $H_0: \mu = 99.2$ against $H_1: \mu \neq 99.2$.

(*vi*) Since $z = 1.792 > 1.645 = z_{0.95}$, we reject $H_0: \mu \leq 99.2$ in favour of $H_1: \mu > 99.2$.

(*vii*) Since $z = -1.428 < -1.282 = z_{0.10}$, we reject H_0: $\mu \geq 86.2$ in favour of H_1: $\mu < 86.2$.

(*viii*) Since $z_{0.005} = -2.576 < z = -2.5 < 2.576 = z_{0.995}$, we do not reject H_0: $\mu = 7.0$ against H_1: $\mu \neq 7.0$. }

3. (*a*) A random sample of size 36 is taken from a normal population with known variance $\sigma^2 = 25$. If the mean of the sample is $\bar{x} = 42.6$, test the null hypothesis $\mu \geq 45$ against the alternative hypothesis $\mu < 45$ with $\alpha = 0.05$ (α is the probability of committing Type-I error).

(Since $z = -2.88 < -1.645 = z_{0.05}$, we reject H_0: $\mu \geq 45$ in favour of H_1: $\mu < 45$)

(*b*) Ten dry cells were taken from store and voltage tests gave the following results: 1.52, 1.53, 1.49, 1.48, 1.47, 1.49, 1.51, 1.50, 1.47, 1.48 volts. The mean voltage of the cells when stored was 1.51V. Assuming the population standard deviation to remain unchanged at 0.02 V, is there reason to believe that the cells have deteriorated.

(Since $z = -2.53 < -1.645 = z_{0.05}$, we reject H_0: $\mu \geq 1.51$ in favour of H_1: $\mu < 1.51$.)

4. (*a*) A sample of 400 male students is found to have a mean height of 67.47 inches. Can it be regarded as a simple random sample from a large population with mean height 67.39 with standard deviation of 1.3 inches?

(Since $z_{0.025} = -1.96 < z = 1.23 < 1.96 = z_{0.975}$, we do not reject H_0: $\mu = 67.39$ against H_1: $\mu \neq 67.39$ at $\alpha = 0.05$)

(*b*) The mean lifetime of electric light bulbs produced by a company has in the past been 1120 hours with a standard deviation of 125 hours. A sample of 8 electric bulbs recently chosen from a supply of newly manufactured bulbs showed a mean lifetime of 1070 hours. Test the hypothesis that mean lifetime of the bulbs has not changed using a level of significance of 0.05.

(Since $z_{0.025} = -1.96 < z = -1.13 < 1.96 = z_{0.975}$, we do not reject H_0: $\mu = 1120$ against H_1: $\mu \neq 1120$)

5. (*a*) Suppose that the mean μ of a random variable X is unknown but the variance of X is known to be 144. Should we reject the null hypothesis H_0: $\mu = 15$ in favour of an alternative hypothesis H_1: $\mu \neq 15$ at a level of significance of $\alpha = 0.05$, if a random sample of 64 observations yields a mean $\bar{x} = 12$? What are the 95% confidence limits for μ.

(Since $z = -2 < -1.960 = z_{0.025}$, we reject H_0: $\mu = 15$ in favour of H_1: $\mu \neq 15$; $9.06 < \mu < 14.94$)

(*b*) In a syrup filling factory the mean weight per bottle is claimed to be $\mu = 16$. A random sample of 100 bottles is taken and a test statistic used is $\bar{X}$, the mean weight of the sample. For a 0.05 level of significance, find the critical values for the test statistic and

formulate the decision rule. Assume the weights to be normally distributed with $\sigma^2 = 2.56$.
(Reject H_0 if $\bar{X} < 15.68$ or $\bar{X} > 16.32$)

6. (*a*) A random sample of 36 drinks from a soft-drink machine has an average content 7.6 ounces with a standard deviation of 0.48 ounces. Test the hypothesis $\mu \leq 7.5$ ounces against the alternative hypothesis $\mu > 7.5$ at the 0.05 level of significance.
(Since $z = 1.25 < 1.645 = z_{0.95}$, we do not reject H_0: $\mu \leq 7.5$ against H_1: $\mu > 7.5$)

(*b*) It is claimed that an automobile is driven on the average at most 20000 kilometres per year. To test this claim, a random sample of 100 automobile owners are asked to keep a record of the kilometres they travel. Would you agree with claim if the random sample showed an average of 21500 kilometres and a standard deviation of 3900 kilometres? Use a 0.01 level of significance.
(Since $z = 3.846 > 2.326 = z_{0.990}$, we reject H_0: $\mu \leq 20000$ in favour of H_1: $\mu > 20000$)

7. (*a*) A random sample of 100 recorded deaths in the United States during the past year showed and average lifespan of 71.8 years with a standard deviation of 8.9 years. Does this seem to indicate that the average lifespan today is greater than 70 years? Use a 0.05 level of significance.
(Since $z = 2.02 > 1.645 = z_{0.95}$, we reject H_0: $\mu \leq 70$ in favour of H_1: $\mu > 70$)

(*b*) It is claimed that an automobile is driven on the average no more than 12000 miles per year. To test this claim, a random sample of 100 automobile owners are asked to keep a record of the miles they travel. Would you agree with the claim if the random sample showed an average of 12500 miles and a standard deviation of 2400 miles?
(Since $z = 2.083 > 1.645$, we reject H_0: $\mu \leq 12000$ in favour of H_1: $\mu > 12000$ at $\alpha = 0.05$)

8. (*a*) Given the following information. What is your conclusion in testing each of the indicated null hypothesis? Assume the populations are normal.

Sample	Sample size	Sample mean	Estimate of variance from sample	Hypotheses		level of significance
				H_0	H_1	
(*i*)	16	11	81	$\mu \leq 10$	$\mu > 10$	0.01
(*ii*)	25	8	64	$\mu \geq 10$	$\mu < 10$	0.01
(*iii*)	25	9	49	$\mu = 10$	$\mu \neq 10$	0.02
Sample	n	$\bar{x}$	$\sum(x_i - \bar{x})^2$	Hypotheses		Level of significance
				H_0	H_1	
(*iv*)	12	24.9	12.3	$-\mu \leq 24.1$	$\mu > 24.1$	2.5%
(*v*)	17	35.6	1471.8	$\mu = 40$	$\mu \neq 40$	5%
(*vi*)	6	1505.8	50.8	$\mu \leq 1503$	$\mu > 1503$	5%
(*vii*)	10	129.8	97.6	$\mu \geq 133.0$	$\mu < 133.0$	1%

{ (*i*) Since $t = 0.44 < 2.602 = t_{15;\,0.99}$, we do not reject H_0.

(*ii*) Since $t = -1.25 > -2.492 = t_{24;\,0.01}$, we do not reject H_0.

(*iii*) Since $t_{24;\,0.01} = -2.492 < t = -0.714 < 2.492 = t_{24;\,0.99}$, we do not reject H_0.

(*iv*) Since $t = 2.622 > 2.201 = t_{11;\,0.975}$, we reject H_0: $\mu \leq 24.1$ in favour of H_1: $\mu > 24.1$.

(*v*) Since $t_{16;\,0.25} = -2.120 < t = -1.892 < 2.120 = t_{16;\,0.975}$ we do not reject H_0: $\mu = 40$ against H_1: $\mu \neq 40$.

(*vi*) Since $t = 2.152 > 2.015 = t_{5;\,0.95}$, we reject H_0: $\mu \leq 1503$ in favour of H_1: $\mu > 1503$.

(*vii*) Since $t = -3.073 < -2.821 = t_{9;\,0.01}$, we reject H_0: $\mu \geq 133.0$ in favour of H_1: $\mu < 133.0$ }.

(*b*) A random sample of size n is drawn from normal population with mean 5 and variance σ^2.

(*i*) If $n = 25$, $\bar{x} = 3$ and $\hat{s} = 2$, what is t?

(*ii*) If $n = 9$, $\bar{x} = 2$ and $t = -2$, what is $\hat{s}$?

(*iii*) If $n = 25$, $\hat{s} = 10$ and $t = 2$, what is $\bar{x}$

(*iv*) If $\hat{s} = 15$, $\bar{x} = 14$ and $t = 3$, what is n?

($t = -5$, $\hat{s} = 4.5$, $\bar{x} = 9$, $n = 25$)

9. (*a*) Injection of a certain type of hormone into hens is said to increase the mean weight of eggs by 0.3 ounces. A sample of 30 eggs has an arithmetic mean 0.4 ounces above the pre-injection mean and a value of $\hat{s}$ equal to 0.20. Is this enough reason to accept the statement that the mean increase is more than 0.3 ounces?

(Since $t = 2.74 > 1.699 = t_{29;\,0.95}$, we reject H_0: $\mu \leq 0.3$ in favour of H_1: $\mu > 0.3$ at $\alpha = 0.05$.)

(*b*) A random sample of 25 hens from a normal population showed that the average laying is 272 eggs per year with a variance of 625 eggs. The company claimed that the average laying is at least 285 eggs per year. Test the claim of the company at $\alpha = 0.05$.

(Since $t = -2.6 < -1.711 = t_{24;\,0.05}$, we reject H_0: $\mu \geq 285$ in favour of H_1: $\mu < 285$).

10. (*a*) A producer of a certain make of flashlight dry cell batteries claims that its output has a mean life of 750 minutes. A random sample of 15 such batteries has been tested and a sample mean of 745 minutes and a sample standard deviation of 24 minutes have been

obtained. Verify that these results are consistent with the null hypothesis $\mu \geq 750$ against $\mu < 750$ at $\alpha = 0.01$.
(Since $t = -0.807 > -2.624 = t_{14;\,0.01}$, we do not reject H_0: $\mu \geq 750$ against H_1: $\mu < 750$.)

(*b*) Ten individuals are chosen at random from a normal population and the heights are found to be 63, 63, 66, 67, 67, 69, 70, 70, 71 and 71inches. In the light of these data, discuss the suggestion that the mean height in the population is 66 inches.
(Since $t_{9;\,0.025} = -2.262 < t = 1.78 < 2.262 = t_{9;\,0.975}$, we do not reject H_0: $\mu = 66$ against H_1: $\mu \neq 66$ at $\alpha = 0.05$.)

11. (*a*) Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 15.90 *oz* and the sum of squared deviations from this mean is 0.276 (oz)^2. Does the sample mean differ significantly from intended weight of 16 *oz*.?
(Since $t_{9;\,0.025} = -2.262 \leq t = -1.806 \leq 2.262 = t_{9;\,0.975}$, we do not reject H_0: $\mu = 16$ against H_1: $\mu \neq 16$ at $\alpha = 0.05$.)

(*b*) In the past a machine has produced washers having thickness of 0.050 inches. To determine whether the machine is in proper working order, a sample of 10 washers is chosen for which the mean thickness is 0.053 inches and the standard deviation is 0.003 inches. Test the hypothesis that the machine is in proper working order using a level of significance of 0.05.
(Since $t = 3.16 > 2.262 = t_{9;\,0.975}$, we reject H_0: $\mu = 0.050$ in favour of H_1: $\mu \neq 0.050$.)

12. (*a*) A random sample of 16 values from a normal population showed a mean of 41.5 inches and a sum of squares of deviations from this mean equal to 135 (inches)^2. Show that the assumption of a mean of 43.5 inches for the population is not reasonable and that the 95% confidence limits for this mean are 39.9 and 43.1 inches.
(Since $t = -2.667 < -2.131 = t_{15;\,0.025}$, we reject H_0: $\mu = 43.5$ in favour of H_1: $\mu \neq 43.5$ at $\alpha = 0.05$.)

(*b*) A random sample of nine from the men of a large city gave a mean height of 68 inches, and the unbiased estimate of the population variance from sample, $\hat{s}^2$ was 4.5 (inches)^2. Are these data consistent with the assumption of a mean height of 68.5 inches for the men of the city.?
(Since $t_{8;\,0.025} = -2.306 < t = -0.708 < 2.306 = t_{8;\,0.975}$, so we do not reject H_0: $\mu = 68.5$ against H_1: $\mu \neq 68.5$ at $\alpha = 0.05$).

◆

13.3 TEST OF HYPOTHESIS ABOUT A POPULATION PROPORTION, π

Tests concerning proportions are based on frequency or count data, which are outcomes of experiments such as the number of defective items in a production line, the number of errors made in typing a complex mathematical manuscript, and so forth. In this section we shall present the tests based on count data, where the test concerns the parameter π of the Bernoulli distribution for both small and large sample sizes. The statistic for testing hypothesis concerning proportions is the number of successes X or the proportion of successes P in n independently repeated Bernoulli trials. For small n, the tests require the use of binomial probabilities. For large n, the normal approximation to the binomial, using the Z statistic, is appropriate, the Z^2 statistic has χ^2-distribution.

13.3.1 Forms of Hypothesis. Let π_0 be the hypothesised value of the population proportion. The three possible null hypotheses and their corresponding alternative hypotheses, are:

1. $H_0: \pi \geq \pi_0$ against $H_1: \pi < \pi_0$
2. $H_0: \pi \leq \pi_0$ against $H_1: \pi > \pi_0$
3. $H_0: \pi = \pi_0$ against $H_1: \pi \neq \pi_0$

In each case, the test will be made by obtaining a simple random sample of size n and counting the number of successes X in the sample and computing the sample proportion P. Then P will be used in computing a test statistic. Depending upon the calculated value of the test statistic, H_0 will be accepted or rejected.

13.3.2 Test based on Normal Approximation. We know that for a Bernoulli distribution with proportion of success π_0, the sampling distribution of the number of successes X in a sample of size n is a binomial distribution with parameters n and π_0. The sampling distribution of the sample proportion $P = X/n$ is also a binomial distribution. The mean of the sampling distribution of P is π_0 and that $\pi_0(1-\pi_0)/n$ is the variance of the sampling distribution. Consequently, the test statistic is

$$Z = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

which is approximately standard normal for large sample because of the normal approximation to the binomial. Or multiplying the numerator and denominator by n, we obtain the test statistic.

$$Z = \frac{X - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

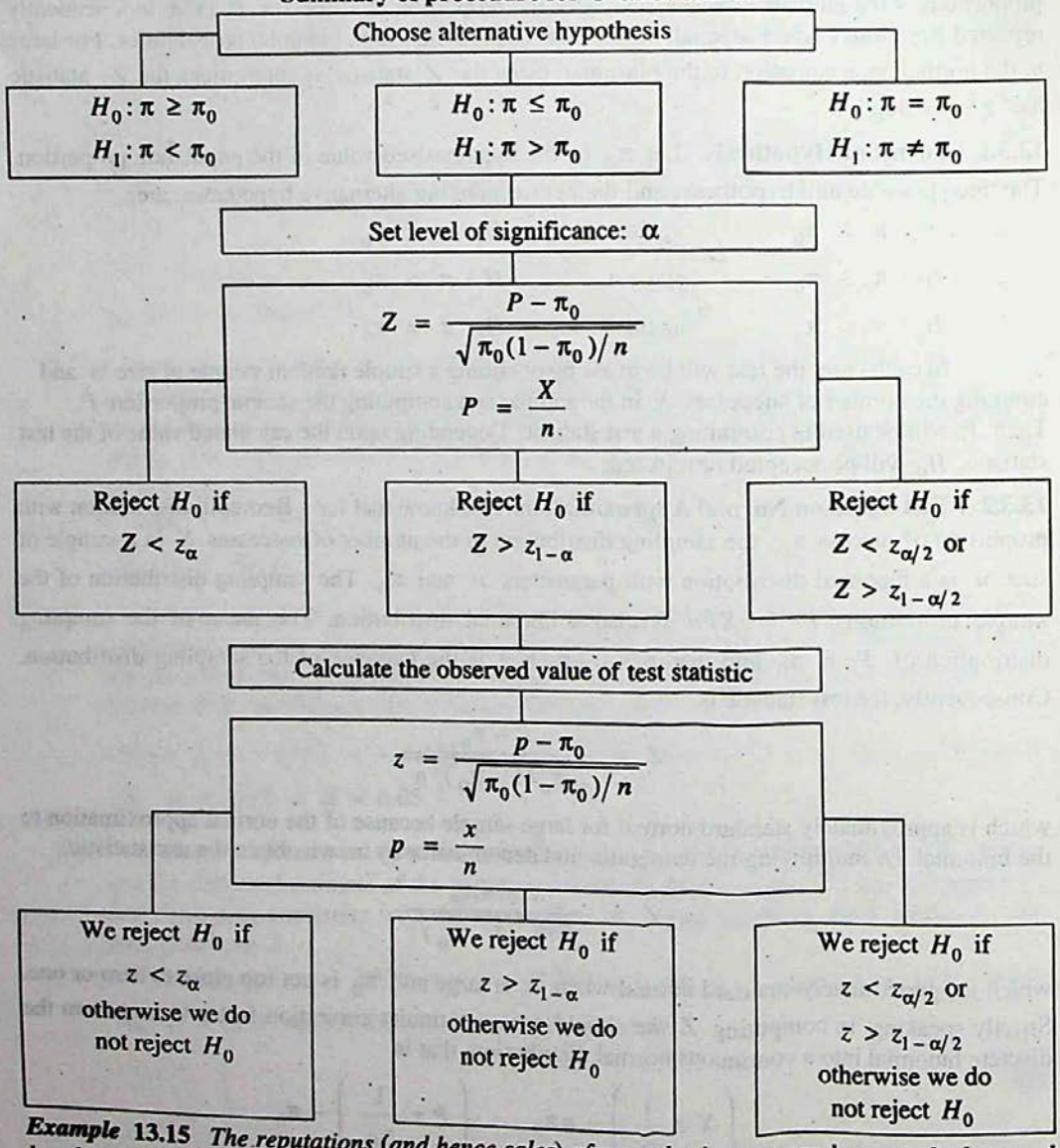
which is approximately standard normal when n is large and π_0 is not too close to zero or one. Strictly speaking, in computing Z we should use a continuity correction factor to transform the discrete binomial into a continuous normal distribution, that is

$$Z = \frac{\left(X \pm \frac{1}{2}\right) - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\left(P \pm \frac{1}{2n}\right) - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

We should use a plus sign (+) when $X < n\,\pi_0$ or $P < \pi_0$ and a minus sign (−) when $X > n\,\pi_0$ or $P > \pi_0$.

The summary of the procedure of testing a hypothesis about the proportion of successes in a population, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing population proportion:
Summary of procedure for 3 alternatives



***Example* 13.15** *The reputations (and hence sales) of many businesses can be severely damaged by shipments of manufactured items that contain an unusually large percentage of defective items. A manufacturer of flashbulbs of cameras may want to be reasonable certain that less than*

5% of its bulbs are defective. Suppose 300 bulbs are randomly selected from a very large shipment, each is tested, and 10 defective bulbs are found. Does this provide sufficient evidence for the manufacture to conclude that the fraction defective in the entire shipment is less than 0.05. Use $\alpha = 0.01$.

Solution. We have $n = 300, \quad x = 10, \quad p = \frac{x}{n} = \frac{10}{300} = 0.033$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the proportion of defectives π is less than 0.05. Consequently, we will test the null hypothesis that $\pi \geq 0.05$ against the alternative hypothesis that $\pi < 0.05$.

The elements of the one-sided left tail test of hypothesis are:

Null hypothesis $H_0 : \pi \geq 0.05$

Alternative hypothesis $H_1 : \pi < 0.05$

Level of significance: $\alpha = 0.01$

Test statistic: $Z = \dfrac{P - \pi_0}{\sqrt{\pi_0 (1 - \pi_0)/n}}$ follows an approximate standard normal distribution under H_0

Critical value: $z_\alpha = z_{0.01} = -2.326$ { From Table 10 (*b*) }

Critical region: $Z < -2.326$

Decision rule: Reject H_0 if $Z < -2.326$, otherwise do not reject H_0.

Observed value: $z = \dfrac{p - \pi_0}{\sqrt{\pi_0 (1 - \pi_0)/n}} = \dfrac{0.033 - 0.05}{\sqrt{0.05(1 - 0.05)/300}} = -1.351$

Conclusion: Since $z = -1.351 > -2.326$, we do not reject H_0. The manufacturer cannot conclude with 99% confidence that the shipment contains fewer than 5% defective bulbs.

***Example* 13.16** *It is known that approximately 10% smokers prefer cigarette brand A. After a promotional campaign in a given sales region, a sample of 200 cigarette smokers were interviewed to determine the effectiveness of the campaign. The results of the survey showed that 26 people expressed a preference of brand A. Do these data present a sufficient evidence to indicate an increase in the acceptance of brand A in the region. Use* $\alpha = 0.05$.

Solution. We have $n = 200, \quad x = 26, \quad p = \frac{x}{n} = \frac{26}{200} = 0.13$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the proportion of smokers π is greater than 0.10. Consequently, we will test the null hypothesis that $\pi \leq 0.10$ against the alternative hypothesis that $\pi > 0.10$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0 : \pi \leq 0.10$

Alternative hypothesis $H_1 : \pi > 0.10$

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $Z = \dfrac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$ follows an approximate standard normal distribution under H_0

Critical value: $z_{1-\alpha} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

Critical region: $Z > 1.645$

Decision rule: Reject H_0 if $Z > 1.645$, otherwise do not reject H_0.

Observed value: $z = \dfrac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \dfrac{0.13 - 0.10}{\sqrt{0.10(1-0.10)/200}} = 1.414$

Conclusion: Since $z = 1.414 < 1.645$, so we do not reject H_0. We cannot conclude with 95% confidence that the promotional campaign is effective.

***Example* 13.17** *A supplier of components to a motor industry makes a particular product which sometimes fails immediately it is used. He controls his manufacturing process so that the proportion of faulty products is supposed to be only* 4%. *Out of* 500 *supplied in one batch* 28 *prove to be faulty. Has the process gone out of control to produce too many faulty products? Test as* $\alpha = 0.05$ *applying continuity correction.*

Solution. We have $n = 500, \quad x = 28$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the proportion of faulty products π is greater than 0.04. Consequently, we will test the null hypothesis that $\pi \leq 0.04$ against the alternative hypothesis that $\pi > 0.04$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \pi \leq 0.04$

Alternative hypothesis $H_1: \pi > 0.04$

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $Z = \dfrac{(X \pm 0.5) - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$ follows an approximate standard normal distribution under H_0

Critical value: $z_{1-\alpha} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

Critical region: $Z > 1.645$

Decision rule: Reject H_0 if $Z > 1.645$, otherwise do not reject H_0.

Observed value: $z = \dfrac{(x \pm 0.5) - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \dfrac{(28 - 0.5) - 500(0.04)}{\sqrt{500(0.04)(1-0.04)}} = 1.71$

Conclusion: Since $z = 1.71 > 1.645$, we reject H_0 and conclude that the process is out of control.

***Example* 13.18** *The records of a certain hospital showed the birth of 723 males and 617 females in a certain week. Do these figures conform to the hypothesis that the sexes are born in equal proportions? Use a* $= 0.02$.

Solution. We have $x = 723, \quad n - x = 617, \quad n = 1340$

$$p = \frac{x}{n} = \frac{723}{1340} = 0.54$$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the proportion of male births π is not equal to 0.5. Consequently, we will test the null hypothesis that $\pi = 0.5$ against the alternative hypothesis that $\pi \neq 0.5$.

The elements of the two-sided test of hypothesis are:

Null hypothesis $H_0 : \pi = 0.5$

Alternative hypothesis $H_1 : \pi \neq 0.5$

Level of significance: $\alpha = 0.02 \Rightarrow \alpha/2 = 0.01 \Rightarrow 1 - \alpha/2 = 0.99$

Test statistic: $Z = \dfrac{P - \pi_0}{\sqrt{\pi_0 (1 - \pi_0)/n}}$ follows an approximate standard normal distribution under H_0

Critical values: $z_{\alpha/2} = z_{0.01} = -2.326,$

$z_{1-\alpha/2} = z_{0.99} = 2.326$ { From Table 10 (*b*) }

Critical region: $Z < -2.326$ or $Z > 2.326$

Decision rule: Reject H_0 if $Z < -2.326$ or $Z > 2.326$, otherwise do not reject H_0

Observed value: $z = \dfrac{p - \pi_0}{\sqrt{\pi_0 (1 - \pi_0)/n}} = \dfrac{0.54 - 0.5}{\sqrt{0.5(1 - 0.5)/1340}} = 2.928$.

Conclusion: Since $z = 2.928 > 2.326$, we reject H_0 and conclude that the proportion of male births is not equal to 0.5.

Exercise 13.3

1. (*a*) For each of the following sets of data, carry out a significance test for the hypotheses stated.

Sample	Number in sample	Number of successes	Hypotheses		Level of significance
			H_0	H_1	
(*i*)	50	45	$\pi \leq 0.8$	$\pi > 0.8$	5%
(*ii*)	60	42	$\pi = 0.55$	$\pi \neq 0.55$	2%
(*iii*)	120	21	$\pi = 1/4$	$\pi \neq 1/4$	5%
(*iv*)	300	213	$\pi = 0.65$	$\pi \neq 0.65$	1%
(*v*)	90	56	$\pi \geq 0.76$	$\pi < 0.76$	1%

{ (*i*) Since $z = 1.768 > 1.645 = z_{0.95}$, we reject $H_0 : \pi \leq 0.8$ in favour of $H_1 : \pi > 0.8$.

(*ii*) Since $z = 2.335 > 2.326 = z_{0.99}$, we reject $H_0 : \pi = 0.55$ in favour of $H_1 : \pi \neq 0.55$.

(*iii*) Since $z_{0.025} = -1.96 < z = 1.897 < 1.96 = z_{0.975}$, we do not reject $H_0: \pi = 1/4$ against $H_1: \pi \neq 1/4$.

(*iv*) Since $z_{0.005} = -2.576 < z = 2.179 < 2.576 = z_{0.995}$, we do not reject $H_0: \pi = 0.65$ against $H_1: \pi \neq 0.65$.

(*v*) Since $z = -3.060 < -2.326 = z_{0.01}$, we reject $H_0: \pi \geq 0.76$ in favour of $H_1: \pi < 0.76$. }

(*b*) A basket ball player has hit on 60% of his shots from the floor. If on the next 100 shots he makes 70 baskets, would you say that his shooting has improved. (Use a 0.05 level of significance).

(Since $z = 2.041 > 1.645 = z_{0.95}$, we reject $H_0: \pi \leq 0.60$ in favour of $H_1: \pi > 0.60$; Yes)

2. (*a*) In a poll of 10000 voters selected at a random from all the voters in a certain district, it is found that 5180 voters are in favour of a particular candidate. Test the null hypotheses that the proportion of all the voters in the district, who favour the candidate is equal to or less than 50% against the alternative that it is greater than 50%. Use a 0.05 level of significance.

(Since $z = 3.6 > 1.645 = z_{0.95}$, we reject $H_0: \pi \leq 0.5$ in favour of $H_1: \pi > 0.5$)

(*b*) A retailer places an order for 400 recapped automobile tires with a supplier who claims that no more than 5% of his output is ever returned unsatisfactory. In time, 31 of the 400 tires are unsatisfactory. Should the retailer continue to trust his supplier's word as to the rate of returns? Use 5% level of significance.

(Since $z = 2.524 > 1.645 = z_{0.95}$, we reject $H_0: \pi \leq 0.05$ in favour of $H_1: \pi > 0.05$; No)

(*c*) A commonly prescribed drug on a market for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension showed that 70 got relief. Is this sufficient evidence that the new drug is superior to the one commonly prescribed? Use 5% level of significance.

(Since $z = 2.041 > 1.645 = z_{0.95}$, we reject $H_0: \pi \leq 0.60$ in favour of $H_1: \pi > 0.60$; Yes)

3. (*a*) An electrical company claimed that at least 95% of the parts which they supplied on a government contract confirmed to specification. A sample of 400 parts was tested, and 355 meet specification. Can we accept the company's claim at a 0.05 level of significance?

(Since $z = -5.735 < -1.645 = z_{0.05}$, we reject $H_0: \pi \geq 0.95$ in favour of $H_1: \pi < 0.95$; No)

(*b*) An electric company claimed that at least 85% of the parts which they supplied conformed to specifications. A sample of 400 parts was tested and 75 did not meet specifications. Can we accept the company's claim at 0.05 level of significance?

(Since $z = -2.100 < -1.645 = z_{0.05}$, we reject $H_0: \pi \geq 0.85$ in favour of $H_1: \pi < 0.85$; No)

(*c*) The manufacturer of a patent medicine claimed that it was at least 90% effective in relieving an allergy for a period of 8 hours. In a sample of 200 people, who had the allergy, the medicine provided relief for 160 people. Determine whether the manufacturer's claim is legitimate.

(Since $z = -4.714 < -1.645 = z_{0.05}$, we reject $H_0: \pi \geq 0.90$ in favour of $H_1: \pi < 0.90$; No)

4. (*a*) A coin is tossed 100 times and 38 heads are obtained. Is there evidence, at the 2% level that the coin is biased in favour of tails?

(Since $z = -2.4 < -2.05 = z_{0.02}$, we reject $H_0: \pi \geq 0.5$ in favour of $H_1: \pi < 0.5$; Yes)

(*b*) A coin is tossed 400 times and it turns up heads 216 times. Discuss whether the coin may be an unbiased one.

(Since $z_{0.25} = -1.960 < z = 1.6 < 1.960 = z_{0.975}$, we do not reject $H_0: \pi = 0.5$ against $H_1: \pi \neq 0.5$)

5. (*a*) In a random sample of 1000 houses in a certain city, 618 own colour TV sets. Is this sufficient evidence to conclude that 2/3 of the houses in this city have colour TV sets? Use $\alpha = 0.02$.

(Since $z = -3.288 < -2.326 = z_{0.01}$, we reject $H_0: \pi = 2/3$ in favour of $H_1: \pi \neq 2/3$; No)

(*b*) The sex distribution of 98 births reported in a newspaper was 52 boys and 46 girls. Is this consistent with an equal sex division in the population? Use 5% level of significance.

(Since $z_{0.25} = -1.960 < z = 0.594 < 1.960 = z_{0.975}$, we do not reject $H_0: \pi = 0.5$ against $H_1: \pi \neq 0.5$.)

6. (*a*) The reputations (and hence sales) of many businesses can be severely damaged by shipments of manufactured items that contain an unusually large percentage of defective items. A manufacturer of flashbulbs of cameras may want to be reasonable certain that less than 5% of its bulbs are defective. Suppose 300 bulbs are randomly selected from a very large shipment, each is tested, and 290 good bulbs are found. Does this provide suficient evidence for the manufacturer to conclude that the fraction defective in the entire shipment is less than 0.05? Use $\alpha = 0.01$.

(Since $z = -1.351 > -2.326 = z_{0.01}$, we do not reject H_0. The manufacturer cannot conclude with 99% confidence that the shipment contains fewer than 5% defective bulbs)

(*b*) A supplier of components to a motor industry makes a particular product which sometimes fails immediately it is used. He controls his manufacturing process so that the proportion of faulty products is supposed to be only 4%. Out of 500 supplied in one batch 28 prove to be faulty. Has the process gone out of control to produce too many faculty products? Test at $\alpha = 0.05$ applying continuity correction.

(Since $z = 1.71 > 1.645 = z_{0.95}$, we reject $H_0: \pi \leq 0.04$ in favour of $H_1: \pi > 0.04$ and conclude that the process is out of control)

◆

13.4 TEST OF HYPOTHESES ABOUT THE DIFFERENCE BETWEEN TWO POPULATION MEANS — *INDEPENDENT SAMPLES*

There are many problems where we are interested in hypotheses concerning about the differences between the means of two populations. For instance, we may wish to decide upon the basis of suitable samples whether a new fertilizer is more effective than an existing fertilizer, or whether a newly introduced product is more reliable than an existing product. Specifically, within the frame work of statistical language, we are interested in making inferences about the parameter $\mu_1 - \mu_2$. A test of hypothesis must be based on assumptions regarding the structure of the underlying distributions. Two independent random samples must be taken — one from each of the two populations of interest.

13.4 .1 Forms of Hypothesis. We are interested in tests about the parameter $\mu_1 - \mu_2$. Let δ_0 be a hypothesized value of the difference between two population means the three possible null hypotheses about the difference between two population means, and their corresponding alternative hypotheses, are:

1. $H_0: \mu_1 - \mu_2 \geq \delta_0$ against $H_1: \mu_1 - \mu_2 < \delta_0$
2. $H_0: \mu_1 - \mu_2 \leq \delta_0$ against $H_1: \mu_1 - \mu_2 > \delta_0$
3. $H_0: \mu_1 - \mu_2 = \delta_0$ against $H_1: \mu_1 - \mu_2 \neq \delta_0$

13.4.2 Independent Samples: ***Normal populations, known variances, any sample sizes.*** Suppose that we have two independent random samples of sizes n_1 and n_2 from two normal populations having, respectively, means μ_1 and μ_2 and known variances σ_1^2 and σ_2^2. We known that $\overline{X}_1$ is distributed as $N(\mu_1, \sigma_1^2/n_1)$ and that $\overline{X}_2$ is distributed as $N(\mu_2, \sigma_2^2/n_2)$, and that $\overline{X}_1$ is independent of $\overline{X}_2$. The statistic $\overline{X}_1 - \overline{X}_2$, the difference between two independent normal variables, is also normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Consequently the random variable

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution. Thus the appropriate test statistic is

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

13.4.3 Independent Samples: ***Normal populations, same unknown variance, small samples.*** Suppose that we have two independent random samples of sizes n_1 and n_2 ($n_1 < 30$ and $n_2 < 30$) from two normal populations having, respectively, means μ_1 and μ_2 and unknown

common variance σ^2 (*i. e.*, $\sigma_1^2 = \sigma_2^2 = \sigma^2$). Then $\overline{X}_1$ is normally distributed with mean μ_1 and variance σ^2/n_1 and $\overline{X}_2$ is normally distributed with mean μ_2 and variance σ^2/n_2 and that $\overline{X}_1$ is independent of $\overline{X}_2$.

Therefore, $\overline{X}_1 - \overline{X}_2$ is normally distributed with mean

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Thus the random variable

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a standard normal distribution. Since we assume that the two populations have equal unknown variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown), we will replace the population variance by the sample variance.

The difficulty in making this replacement lies in the fact that we have two estimates of σ^2, $\hat{S}_1^2$ and $\hat{S}_2^2$, since two different samples were collected. Even if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, it is unlikely that the two samples collected will have exactly the same value because of sampling error. But if $\hat{S}_1^2$ and $\hat{S}_2^2$ differ which of these two estimates should be used to estimate the unknown population variance σ^2.

Since we wish to obtain the best estimate available, it would seem reasonable to use an estimator that would pool the information from both samples. Thus, if $\hat{S}_1^2$ and $\hat{S}_2^2$ are the two sample variances (both estimating the variance σ^2 common to both populations), the pooled (weighted arithmetic mean) estimator of σ^2, denoted by S_p^2, is

$$S_p^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

$$= \frac{\Sigma(X_{i1} - \overline{X}_1)^2 + \Sigma(X_{i2} - \overline{X}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{(\Sigma X_{i1}^2 - n_1\overline{X}_1^2) + (\Sigma X_{i2}^2 - n_2\overline{X}_2^2)}{n_1 + n_2 - 2}$$

To obtain the small-samples test statistic for testing $H_0: \mu_1 - \mu_2 = \delta_0$, substitute the pooled estimator of σ^2 into the above formula to obtain

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.

13.4.4 Independent Samples: ***Any populations, large samples.*** When both sample sizes are large (say greater than 30) the assumptions regarding small samples can be greatly relaxed. It is no longer necessary to assume that the parent distributions are normal, because the Central Limit Theorem assures that $\bar{X}_1$ is approximately normally distributed with mean μ_1 and variance σ_1^2/n_1, and that $\bar{X}_2$ is also approximately normally distributed with mean μ_2 and variance σ_2^2/n_2, and that $\bar{X}_1$ is independent of $\bar{X}_2$, then $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

Thus the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has an approximate standard normal distribution.

Because n_1 and n_2 are both large, the approximation remains valid if σ_1^2 and σ_2^2 are replaced by their sample variances $\hat{S}_1^2$ and $\hat{S}_2^2$. The assumption of equal variance is not required in inferences derived from large samples. We can still use Z test with $\hat{S}_1^2$ substituted for σ_1^2 and $\hat{S}_2^2$ substituted for σ_2^2 so long as both samples are large enough for the Central Limit Theorem to be applied.

That is, we use

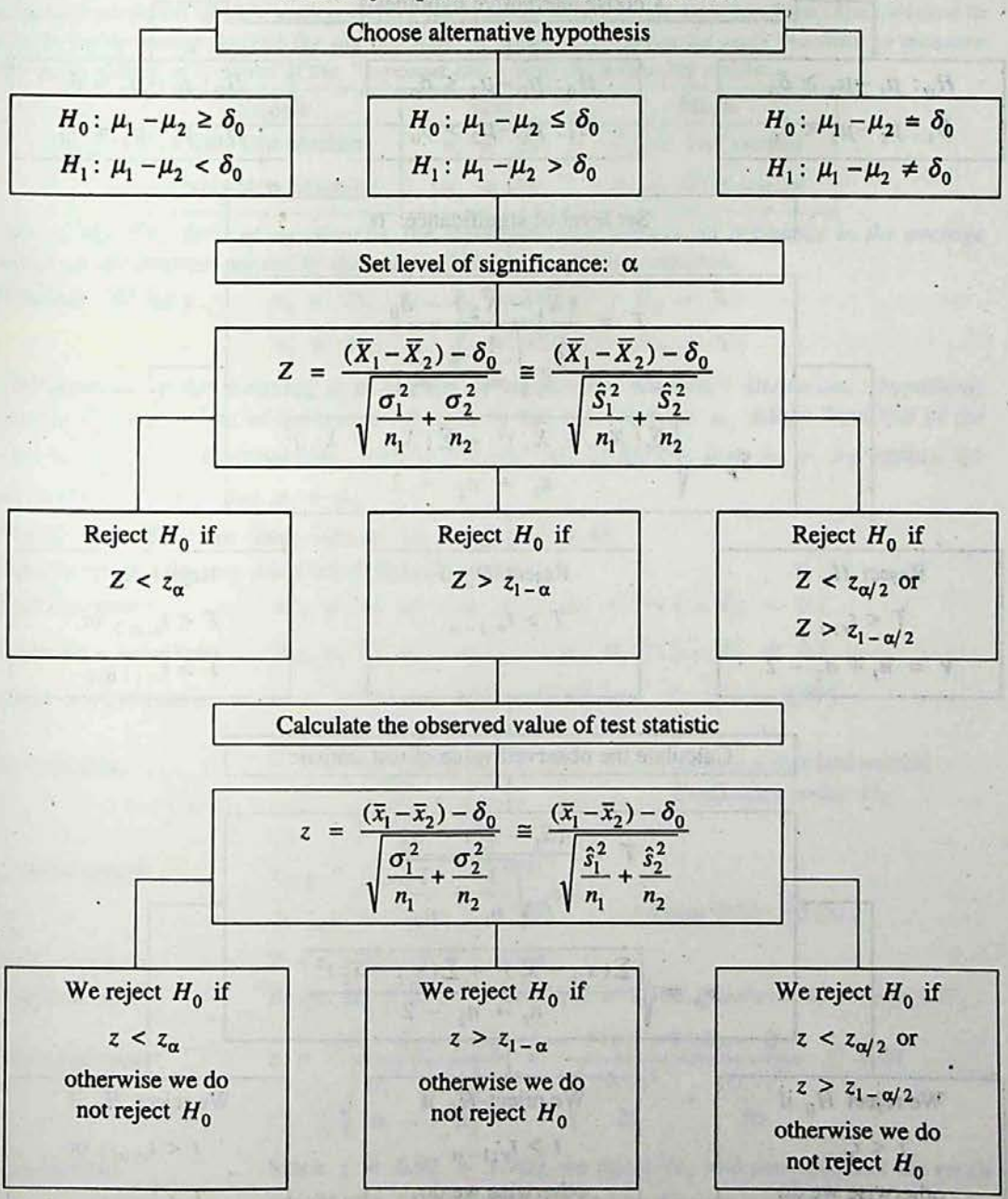
$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

as the test statistic.

The summary of the procedure of testing a hypothesis about the difference between means of two populations whose variances σ_1^2, σ_2^2 are known/unknown, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

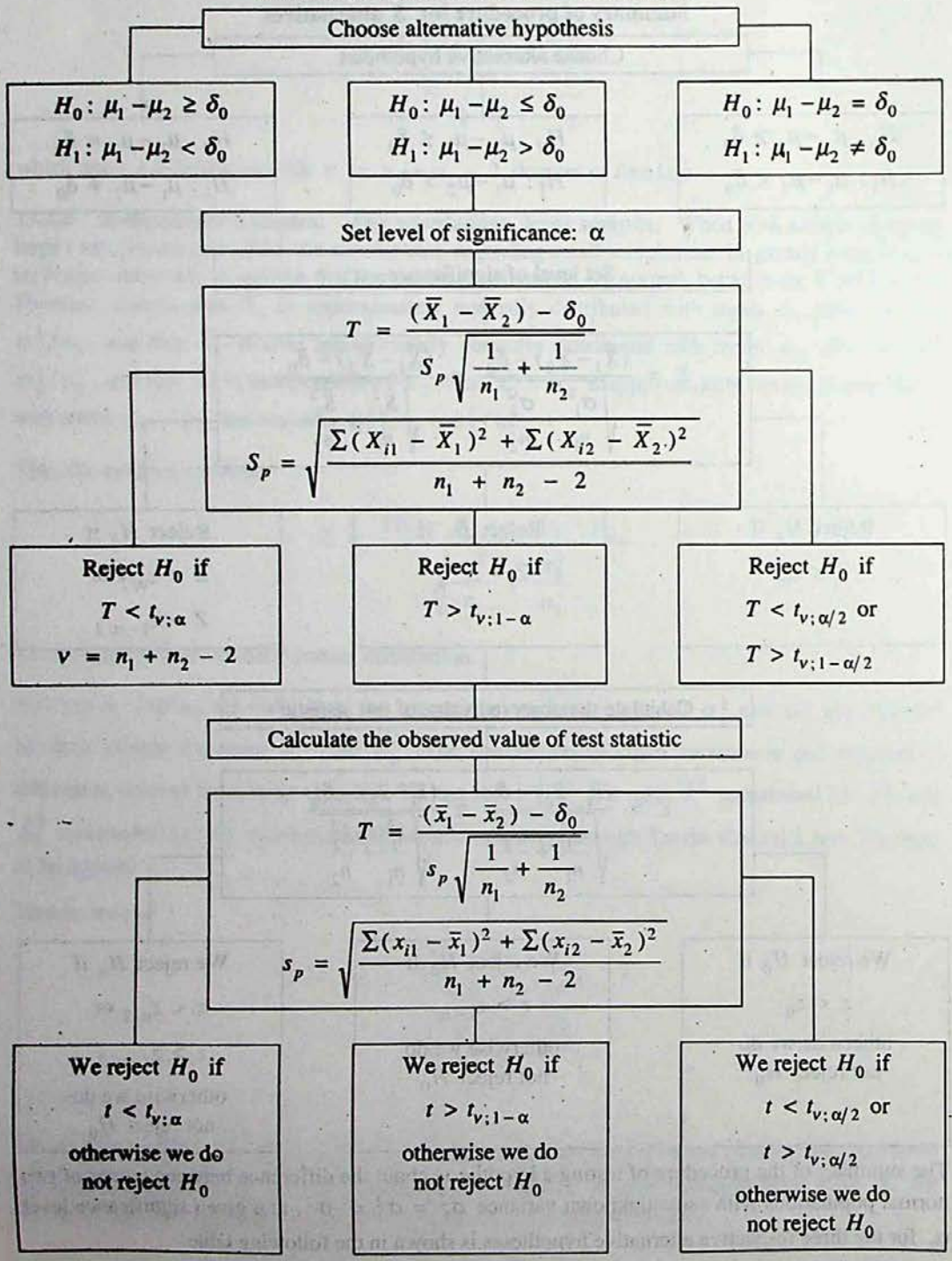
Testing difference between means of two populations, σ_1^2, σ_2^2 known/unknown:

Summary of procedure for 3 alternatives



The summary of the procedure of testing a hypothesis about the difference between means of two normal populations with same unknown variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing difference between means of two normal populations, same unknown variance: Summary of procedure for 3 alternatives



***Example* 13.19** *Apex's current packaging machinery for coffee is known to ground coffee into "1-pound cans" with a standard deviation of 0.6 ounce. Apex is considering using a new packaging machine which is expected to pour coffee into "1-pound cans" more accurately, with a standard deviation of 0.3 ounce. Before deciding to invest in the new machine, Apex wished to test its performance against the old machine. A sample was taken on each machine to measure the mean weight of contents of the "1-pound can", with the following results.*

Sample	Size	Mean
Using Old Machine	$n_1 = 25$	$\bar{x}_1 = 16.7$ ounces
Using New Machine	$n_2 = 36$	$\bar{x}_2 = 15.8$ ounces

Test, at the 5% level of significance, the hypothesis that there is no difference in the average weight of the contents poured by the old machine versus the new machine.

Solution. We have $n_1 = 25, \quad \bar{x}_1 = 16.7, \quad \sigma_1 = 0.6$

$n_2 = 36, \quad \bar{x}_2 = 15.8, \quad \sigma_2 = 0.3$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the average weight of the contents poured by the old machine μ_1 differs from that of the new machine μ_2. Consequently, we will test the null hypothesis that $\mu_1 = \mu_2$ against the alternative hypothesis that $\mu_1 \neq \mu_2$.

The elements of the one-sided right tail test of hypothesis are:

The elements of the two-sided test of hypothesis are:

Null hypothesis $H_0: \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0 \; (i.e.\ \delta_0 = 0)$

Alternative hypothesis $H_1: \mu_1 \neq \mu_2 \Rightarrow \mu_1 - \mu_2 \neq 0 \; (i.e.\ \delta_0 \neq 0)$

Level of significance: $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$

Test statistic: $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ follows a standard normal distribution under H_0

Critical values: $z_{\alpha/2} = z_{0.025} = -1.960,$

$z_{1-\alpha/2} = z_{0.975} = 1.960$ { From Table 10 (*b*) }

Critical region: $Z < -1.960$ or $Z > 1.960$

Decision rule: Reject H_0 if $Z < -1.960$ or $Z > 1.960$, otherwise do not reject H_0.

Observed value: $$z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(16.7 - 15.8) - 0}{\sqrt{\frac{(0.6)^2}{25} + \frac{(0.3)^2}{36}}} = 6.92$$

Conclusion: Since $z = 6.92 > 1.960$, we reject H_0 and conclude that the mean weight of a "1-pound can" filled by the new machine is not the same as that of filled by the old machine.

***Example* 13.20** *The test was given to a group of 100 scouts and to a group of 144 guides. The mean score for the scouts was 27.53 and the mean score for the guides was 26.81.*

Assuming a common population standard deviation of 3.48, test, using a 5% level of significance, whether the scouts performance in the test was better than that of the guides.

Solution. We have $n_1 = 100, \quad \bar{x}_1 = 27.53$

$n_2 = 144, \quad \bar{x}_2 = 26.81$

Common population standard deviation: $\sigma = 3.48$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean score for the scouts μ_1 is greater than that of the guides μ_2. Consequently, we will test the null hypothesis that $\mu_1 \leq \mu_2$ against the alternative hypothesis that $\mu_1 > \mu_2$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu_1 \leq \mu_2 \Rightarrow \mu_1 - \mu_2 \leq 0 \; (i.e. \; \delta_0 \leq 0)$

Alternative hypothesis $H_1: \mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 > 0 \; (i.e. \; \delta_0 > 0)$

Level of significance: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Test statistic: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ follows a standard normal distribution under H_0

Critical value: $z_{1-\alpha} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

Critical region: $Z > 1.645$

Decision rule: Reject H_0 if $Z > 1.645$, otherwise do not reject H_0.

Observed value: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{(27.53 - 26.81) - 0}{3.48\sqrt{\dfrac{1}{100} + \dfrac{1}{144}}} = 1.589$

Conclusion: Since $z = 1.589 < 1.645$, we do not reject H_0 and conclude that there is not sufficient evidence, at 5% level of significance, to show that the performance of the scouts in the test was better than that of the guides.

***Example* 13.21** *The management of a restaurant wants to determine whether a new advertising campaign has increased its mean daily income (gross). The income for 50 business days prior to the campaign's beginning were recorded. After conducting the advertising campaign and allowing a 20 days period for the advertising to take effect, the restaurant management recorded the income for 30 business days. These two samples will allow the management to make an inference about the effect of the advertising campaign on the restaurant's daily income. A summary of the results of the two samples are shown below:*

Sample	*Size*	*Mean*	*Standard deviation*
Before campaign	$n_1 = 50$	$\bar{x}_1 = 1255$	$\hat{s}_1 = 215$
After campaign	$n_2 = 30$	$\bar{x}_2 = 1330$	$\hat{s}_2 = 238$

Do these samples provide sufficient evidence for the management to conclude that the mean income has been increased by the advertising campaign test using $\alpha = 0.05$*?*

Solution. The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean daily income after advertising campaign μ_2 is greater than that of before the campaign μ_1. Consequently, we will test the null hypothesis that $\mu_2 \leq \mu_1$ against the alternative hypothesis that $\mu_2 > \mu_1$.

The elements of the one-sided right tail test of hypothesis are

Null hypothesis $H_0: \mu_2 \leq \mu_1 \Rightarrow \mu_2 - \mu_1 \leq 0 \; (i.e.\ \delta_0 \leq 0)$

Alternative hypothesis $H_1: \mu_2 > \mu_1 \Rightarrow \mu_2 - \mu_1 > 0 \; (i.e.\ \delta_0 > 0)$

Level of significance: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Test statistic: $$Z = \frac{(\bar{X}_2 - \bar{X}_1) - \delta_0}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$ follows a standard normal distribution under H_0

Critical value: $z_{1-\alpha} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

Critical region: $Z > 1.645$

Decision rule: Reject H_0 if $Z > 1.645$, otherwise do not reject H_0

Observed value: $$z = \frac{(\bar{x}_2 - \bar{x}_1) - \delta_0}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} = \frac{(1330 - 1255) - 0}{\sqrt{\frac{(215)^2}{50} + \frac{(238)^2}{30}}} = 1.414$$

Conclusion: Since $z = 1.414 < 1.645$, we do not reject H_0. That is, the samples do not provide sufficient evidence, at the $\alpha = 0.05$ significance level, for the restaurant management to conclude that the advertising campaign has increased the mean daily income.

***Example* 13.22** *A feeding test is conducted on a herd of* 25 *milking cows to compare two diets, one of dewatered alfalfa and the other of field-wilted alfalfa. Dewatered alfalfa has an economic advantage in that its mechanical processing produces a liquid protein-rich by product that can be used to supplement the feed of other animals. A sample of* 12 *cows randomly selected from the herd are fed dewatered alfalfa; the remaining* 13 *cows are fed field-wilted alfalfa. From observations made over a three-week period, the average daily milk production in pounds recorded for each cow is*:

Field-wilted alfalfa	44	44	56	46	47	38	58	53	49	35	46	30	41
Dewatered alfalfa	35	47	55	29	40	39	32	41	42	57	51	39	

Do the data strongly indicate that the milk yield is less with dewatered alfalfa than with field-wilted alfalfa? Test at $\alpha = 0.05$.

Solution. The means of two samples and estimate of common variance are

x_1	44	44	56	46	47	38	58	53	49	35	46	30	41	$\sum x_1 = 587$
x_2	35	47	55	29	40	39	32	41	42	57	51	39		$\sum x_2 = 507$
x_1^2	1936	1936	3136	2116	2209	1444	3364	2809	2401	1225	2116	900	1681	$\sum x_1^2 = 27273$
x_2^2	1225	2209	3025	841	1600	1521	1024	1681	1764	3249	2601	1521		$\sum x_2^2 = 22261$

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{587}{13} = 45.15$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{507}{12} = 42.25$$

$$\Sigma(x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1} = 27273 - \frac{(587)^2}{13} = 767.69$$

$$\Sigma(x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2} = 22261 - \frac{(507)^2}{12} = 840.25$$

$$s_p = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{767.69 + 840.25}{13 + 12 - 2}} = 8.36$$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean milk yield with dewatered alfalfa μ_2 is less than with field-wilted alfalfa μ_1. Consequently, we will test the null hypothesis that $\mu_2 \geq \mu_1$ against the alternative hypothesis that $\mu_2 < \mu_1$.

The elements of the one-sided left tail test of hypothesis are:

Null hypothesis $H_0: \mu_2 \geq \mu_1 \Rightarrow \mu_2 - \mu_1 \geq 0$ (*i.e.* $\delta_0 \geq 0$)

Alternative hypothesis $H_1: \mu_2 < \mu_1 \Rightarrow \mu_2 - \mu_1 < 0$ (*i.e.* $\delta_0 < 0$)

Level of significance: $\alpha = 0.05$

Test statistic: $T = \dfrac{(\bar{X}_2 - \bar{X}_1) - \delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n_1 + n_2 - 2 = 13 + 12 - 2 = 23$

Critical value: $t_{\nu;\alpha} = t_{23;0.05} = -1.714$ (From Table 12)

Critical region: $T < -1.714$

Decision rule: Reject H_0 if $T < -1.714$, otherwise do not reject H_0

Observed value: $t = \dfrac{(\bar{x}_2 - \bar{x}_1) - \delta_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{(42.25 - 45.15) - 0}{8.36\sqrt{\dfrac{1}{13} + \dfrac{1}{12}}} = -0.87$

Conclusion: Since $t = -0.87 > -1.714$, we do not reject $H_0: \mu_2 \geq \mu_1$ against $H_1: \mu_2 < \mu_1$.

***Example* 13.23** *Two random samples taken independently from normal populations with an identical variance yield the following results.*

Sample	Size	Mean	Variance
I	$n_1 = 12$	$\bar{x}_1 = 10$	$\hat{s}_1^2 = 1200$
II	$n_2 = 18$	$\bar{x}_2 = 25$	$\hat{s}_2^2 = 900$

Test the hypotheses that the true difference between the population means is at most 10, *that is* $\mu_2 - \mu_1 \leq 10$, *against the alternative that* $\mu_2 - \mu_1 > 10$ *at 5% level of significance.*

Solution. The estimate of common variance is

$$s_p = \sqrt{\frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(12 - 1)1200 + (18 - 1)900}{12 + 18 - 2}} = 31.9$$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the mean of second population μ_2 is greater than that of first population μ_1 by more than 10 points. Consequently, we will test the null hypothesis that $\mu_2 - \mu_1 \leq 10$ against the alternative hypothesis that $\mu_2 - \mu_1 > 10$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu_2 - \mu_1 \leq 10$ (*i. e.* $\delta_0 \leq 10$)

Alternative hypothesis $H_1: \mu_2 - \mu_1 > 10$ (*i. e.* $\delta_0 > 10$)

Level of significance: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Test statistic: $T = \dfrac{(\bar{X}_2 - \bar{X}_1) - \delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n_1 + n_2 - 2 = 12 + 18 - 2 = 28$

Critical values: $t_{\nu; 1-\alpha} = t_{28; 0.95} = 1.701$ (From Table 12)

Critical region: $T > 1.701$

Decision rule: Reject H_0 if $T > 1.701$, otherwise do not reject H_0.

Observed value: $t = \dfrac{(\bar{x}_2 - \bar{x}_1) - \delta_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{(25 - 10) - 10}{31.9\sqrt{\dfrac{1}{12} + \dfrac{1}{18}}} = 0.421$

Conclusion: Since $t = 0.421 < 1.701 = t_{28; 0.95}$, we do not reject H_0.

Exercise 13.4

1. (*a*) For each of the following sets of data, perform a test to decide whether there is a significant difference between the means, μ_1 and μ_2, of the normal populations from which the samples are drawn.

	n_1	$\sum x_1$	σ_1^2	n_2	$\sum x_2$	σ_2^2	Hypotheses	α
(*i*)	100	4250	30	80	3544	35	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	5%
(*ii*)	20	95	2.3	25	135	2.5	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	2%
(*iii*)	50	1545	6.5	50	1480	7.1	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	1%

	n_1	$\sum x_1$	n_2	$\sum x_2$	Common Population standard deviation σ	Hypotheses	α
(*iv*)	100	12730	100	12410	10.9	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	5%
(*v*)	30	192	45	315	1.25	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	1%
(*vi*)	200	18470	300	27663	0.86	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	10%

	n_1	$\sum x_1$	$\sum(x_1-\bar{x}_1)^2$	n_2	$\sum x_2$	$\sum(x_2-\bar{x}_2)^2$	Hypotheses	α
(*vii*)	40	2128	810	50	2580	772	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	5%
(*viii*)	80	6824	2508	100	8740	3969	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	2%
(*ix*)	65	5369	8886	80	4672	5026	$H_0: \mu_1 - \mu_2 \leq 20$ $H_1: \mu_1 - \mu_2 > 20$	1%

{ (*i*) Since $z = -2.096 < -1.96 = z_{0.025}$, we reject $H_0: \mu_1 = \mu_2$ in favour of $H_1: \mu_1 \neq \mu_2$.

(*ii*) Since $z = -1.402 > -2.054 = z_{0.02}$, we do not reject $H_0: \mu_1 \geq \mu_2$ against $H_1: \mu_1 < \mu_2$.

(*iii*) Since $z = 2.493 > 2.326 = z_{0.99}$, we reject $H_0: \mu_1 \leq \mu_2$ in favour of $H_1: \mu_1 > \mu_2$.

(*iv*) Since $z = 2.076 > 1.645 = z_{0.95}$, we reject $H_0: \mu_1 = \mu_2$ in favour of $H_1: \mu_1 > \mu_2$.

(*v*) Since $z = -2.036 > -2.326 = z_{0.01}$, we do not reject $H_0: \mu_1 \geq \mu_2$ against $H_1: \mu_1 < \mu_2$.

(*vi*) Since $z = 1.783 > 1.645 = z_{0.95}$, we reject $H_0: \mu_1 = \mu_2$ in favour of $H_1: \mu_1 \neq \mu_2$.

(*vii*) Since $z = 1.752 > 1.645 = z_{0.95}$, we reject $H_0: \mu_1 \leq \mu_2$ in favour of $H_1: \mu_1 > \mu_2$.

(*viii*) Since $z_{0.01} = -2.326 < z = -2.351 < 2.326 = z_{0.99}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$.

(*ix*) Since $z = 2.453 > 2.326 = z_{0.99}$, we reject $H_0: \mu_1 - \mu_2 \leq 20$ in favour of $H_1: \mu_1 - \mu_2 > 20$.}

(*b*) A simple sample of heights of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2.56 inches, while a simple sample of heights of 6400 Australian has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that Australians are on the average taller than the Englishmen?

{ Since $z = 15.589 > 1.645 = z_{0.95}$, we reject $H_0: \mu_2 \leq \mu_1$ against $H_1: \mu_2 > \mu_1$ at $\alpha = 0.05$. }

2. (*a*) A random sample of 100 professors in private colleges showed an average monthly salary of Rs. 5000 with a standard deviation of Rs. 200. Another random sample of 150 professors in Govt. Colleges showed an average monthly salary of Rs. 5600 with a standard deviation of Rs. 250. Test the hypotheses that the average salary for the professors teaching in Govt. Colleges does not exceed the average salary for professors teaching in private colleges by more than Rs. 500. Use $\alpha = 0.01$.

{ Since $z = 3.499 > 2.326 = z_{0.99}$, we reject $H_0: \mu_2 - \mu_1 \leq 500$ in favour of $H_1: \mu_2 - \mu_1 > 500$. }

(*b*) A random sample of 80 light bulbs manufactured by company *A* had an average lifetime of 1258 hours with a standard deviation of 94 hours, while a random sample of 60 light bulbs manufactured by company *B* had an average lifetime of 1029 hours with a standard deviation of 68 hours. Because of the high cost of bulbs from company *A*, we are inclined to buy from company *B* unless the bulbs from company *A* will last over 200 hours longer on the average than those from company *B*. Run a test using $\alpha = 0.01$ to determine from whom we should buy our bulbs.

{ Since $z = 2.118 < 2.326 = z_{0.99}$, we do not reject $H_0: \mu_1 - \mu_2 \leq 200$ against $H_1: \mu_1 - \mu_2 > 200$. }

3. (*a*) A farmer claims the average yield of corn of variety *A* exceeds the average yield of variety *B* by at least 12 bushels per acre. To test this claim, 50 acres of each variety are planted and grown under similar conditions. Variety *A* yields on the average 86.7 bushels per acre with a standard deviation of 6.28 bushels per acre, while Variety *B* yields on the average 77.8 bushels per acre with a standard deviation of 5.61 bushels per acre. Test the farmer's claim using a 0.05 level of significance.

{ Since $z = -2.603 < -1.645 = z_{0.05}$, we reject $H_0: \mu_1 - \mu_2 \geq 12$ in favour of $H_1: \mu_1 - \mu_2 < 12$. }

(*b*) An examination was taken to two classes of 40 and 50 students respectively. In the first class, mean grade was 74 with a standard deviation of 8, while in the second class the mean grade was 78 with a standard deviation of 7. Is there a significance

difference between the mean grades at 1% level.

{ Since $z_{0.005} = -2.576 < z = 2.490 < 2.576 = z_{0.995}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$. }

4. (*a*) A random sample of 100 light bulbs manufactured by company A had a mean lifetime of 1230 hours with a standard deviation of 80 hours, while a random sample of 121 light bulbs manufactured by company B had a mean lifetime of 1200 hours with a standard deviation of 66 hours. Are the mean lifetimes of bulbs manufactured by two companies are significantly different?

{ Since $z = 3 > 1.960 = z_{0.975}$, we reject $H_0: \mu_1 = \mu_2$ in favour of $H_1: \mu_1 \neq \mu_2$. }

(*b*) A random sample of 200 villages was taken from Faisalabad District and average population per village was found to be 498 with a standard deviation of 50. Another sample of 200 villages from the same district gave an average population 510 per village with a standard deviation of 40. Is the difference between the average of the two samples statistically significant ?

{ Since $z = -2.650 < -1.960 = z_{0.025}$, we reject $H_0: \mu_1 = \mu_2$ in favour of $H_1: \mu_1 \neq \mu_2$. }

(*c*) The means of simple samples of 500 and 400 are 11.5 and 10.9 respectively. Can the samples be regarded as drawn from a population of standard deviation 5 ?

{ Since $z_{0.025} = -1.960 < z = 1.789 < 1.960 = z_{0.975}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$. }

5. (*a*) Describe the procedure for testing the equality of means of two normal populations for:

(*i*) Large samples,

(*ii*) Small samples

(*b*) For each of the following sets of data, perform a test to decide whether there is a significant difference between the means, μ_1 and μ_2, of the normal populations from which the samples are drawn.

	n_1	Σx_1	$\Sigma(x_1-\bar{x}_1)^2$	n_2	Σx_2	$\Sigma(x_2-\bar{x}_2)^2$	Hypotheses	α
(*i*)	6	171	83	7	164.5	112	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	5%
(*ii*)	5	678.5	562.3	7	971.6	308.6	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	5%
(*iii*)	8	238.4	296	10	206	145	$H_0: \mu_1 - \mu_2 \leq 4$ $H_1: \mu_1 - \mu_2 > 4$	1%
(*iv*)	12	116.16	45.1	18	156.96	72	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	10%

{ (*i*) Since $t = 2.135 > 1.796 = t_{11;\,0.95}$, we reject $H_0: \mu_1 \leq \mu_2$ in favour of $H_1: \mu_1 > \mu_2$.

(*ii*) Since $t_{10;\,0.025} = -2.228 < t = -0.567 < 2.228 = t_{10;\,0.975}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$.

(*iii*) Since $t = 2.088 < 2.583 = t_{16;\,0.99}$, we do not reject $H_0: \mu_1 - \mu_2 \leq 4$ against $H_1: \mu_1 - \mu_2 > 4$.

(*iv*) Since $t_{28;\,0.05} = -1.701 < t = 1.260 < 1.701 = t_{28;\,0.95}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$.}

(*c*) Eight pots, growing three barley plants each, were exposed to a high tension discharge while nine similar pots were enclosed in an earthen wire cage, the number of tillers (shoots) in each pot were as follows:

Caged	17	18	25	27	29	27	23	17	28
Electrified	16	16	20	16	21	17	15	20	

Discuss whether the electrification exercises any real effect on tillering.

{ Since $t = 3.058 > 1.753 = t_{15;\,0.95}$, we reject $H_0: \mu_1 \leq \mu_2$ in favour of $H_1: \mu_1 > \mu_2$ at $\alpha = 0.05$. }

6. (*a*) A random sample of 6 cows of breed *A* had daily milk yields in lb., as 16, 15, 18, 17, 19 and 17 and another random sample of 8 cows of breed *B* had daily milk yields in lb., as 18, 22, 21, 23, 19, 20, 24 and 21. Test if breed *B* is better than breed *A* at $\alpha = 0.05$.

{ Since $t = 4.162 > 1.782 = t_{12;\,0.95}$, we reject $H_0: \mu_2 \leq \mu_1$ in favour of $H_1: \mu_2 > \mu_1$. }

(*b*) The heights of six randomly selected sailors are in inches: 62, 64, 67, 68, 70 and 71. Those of ten randomly selected soldiers are 62, 63, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss in the light of these data that soldiers are on the average taller than sailors. Assume that the heights are normally distributed.

{ Since $t = 0.526 < 1.761 = t_{14;\,0.95}$, we do not reject $H_0: \mu_2 \leq \mu_1$ against $H_1: \mu_2 > \mu_1$ at $\alpha = 0.05$. }

7. (*a*) The weights in grams of 10 male and 10 female juvenile ring-necked pheasants are:

Males	1293	1380	1614	1497	1340	1643	1466	1627	1383	1711
Females	1061	1065	1092	1017	1021	1138	1143	1094	1270	1028

Test the hypothesis of a difference of 350 grams between population means in favour of males against the alternative of a greater difference.

{ Since $t = 1.007 < 1.734 = t_{18;\,0.95}$, we do not reject $H_0: \mu_1 - \mu_2 \leq 350$ against $H_1: \mu_1 - \mu_2 > 350$ at $\alpha = 0.05$. }

(*b*) The following data are the gains in weight, measured in pounds, of babies from birth to age six months. All babies in both groups weighed approximately the same at birth. The babies in sample I were fed formula *A*, and babies in sample II were fed formula *B*.

(Assume that the experimenter has no preconceived notions about which formula might be better).

Sample I	5	7	8	9	6	7	10	8	6
Sample II	9	10	8	6	8	7	9		

Test at the 5% level of significance that the mean of population I equals mean of population II.

{ Since $t_{14;\,0.025} = -2.145 < t = -1.083 < 2.145 = t_{14;\,0.975}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$. }

8. (*a*) From the area planted in one variety of value, 54 plants were selected at random. Of these plants, 15 were "Off-types" and 12 were "Abberrant". The rubber percentages for these plants were:

Off-types	6.21,	5.70,	6.04,	4.47,	5.22,	4.45,	4.84,	5.88
	5.82,	6.09,	5.59,	6.06,	5.59,	6.74,	5.55.	
Aberrants	4.28,	7.71,	6.48,	7.71,	7.37,	7.20,	7.06,	6.40,
	8.93,	5.91,	5.51,	6.36.				

Test the hypothesis of no difference between the two means. Also compute a 95% confidence interval for the difference of two population means.

{ Since $t = -3.120 < -2.060 = t_{25;\,0.025}$, we reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$; $0.383 < \mu_2 - \mu_1 < 1.871$. }

(*b*) The I.Q.'s of 16 students from one area of a city showed a mean of 107 with a standard deviation of 10, while the I.Q.'s of 14 students from another area of the city showed a mean of 115 with a standard deviation of 8. Is there a significant difference between the I.Q.'s of the two groups at (*i*) 0.01 and (*ii*) 0.05 level of significance ?

{ (*i*) Since $t_{28;\,0.005} = -2.763 < t = -2.395 < 2.763 = t_{28;\,0.995}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ at $\alpha = 0.01$.

(*ii*) Since $t = -2.395 < -2.048 = t_{28;\,0.025}$, we reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$; at $\alpha = 0.05$. }

9. (*a*) Means of random samples, each of size 10, from two normal populations with the same standard deviation were found to be 16 and 20 respectively. Further, the sample standard deviations were equal to 5 and 7 respectively. Test the hypotheses that the populations have the same mean, using 0.05 level of significance.

{ Since $t_{18;\,0.025} = -2.101 < t = -1.47 < 2.101 = t_{18;\,0.975}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$. }

(*b*) A random sample of size $n_1 = 10$, selected from a normal population has a mean $\bar{x}_1 = 20$ and a standard deviation $\hat{s}_1 = 5$. A second random sample of size $n_2 = 12$, selected from a different normal population has a mean $\bar{x}_2 = 24$ and a standard deviation $\hat{s}_2 = 6$. If $\mu_1 = 22$ and $\mu_2 = 19$ and σ_1^2 and σ_2^2 are unknown but

approximately equal, test whether there is any reason to doubt that $\mu_1 - \mu_2 = 3$.

{ Since $t = -2.934 < -2.086 = t_{20;\,0.025}$, we reject $H_0: \mu_1 - \mu_2 = 3$ in favour of $H_1: \mu_1 - \mu_2 \neq 3$ at $\alpha = 0.05$. }

10. (*a*) The following values are obtained by two different samples while sampling a normal population with $\mu = 3.0$ and $\sigma = 0.5$. Using these data solve the following problems:

A	2.7	3.9	2.6	2.8	3.2	3.6	2.7	3.0	3.8	3.1
B	2.4	2.8	3.0	3.3	3.7	3.3	3.1	2.7	2.4	3.3

(*i*) Combine the data into one set and test the hypothesis $H_0: \mu = 3.0$ against $H_1: \mu \neq 3.0$.

(*ii*) Assuming that σ is unknown test the hypothesis $H_0: \mu_A \leq \mu_B$ against $H_1: \mu_A > \mu_B$. Use $\alpha = 0.05$ in both cases.

{ (*i*) Since $z_{0.025} = -1.96 < z = 0.626 < 1.96 = z_{0.975}$, we do not reject $H_0: \mu = 3$ against $H_1: \mu_1 \neq 3$. (*ii*) Since $t = 0.694 < 1.734 = t_{18;\,0.95}$, we do not reject $H_0: \mu_A \leq \mu_B$ against $H_1: \mu_A > \mu_B$. }

(*b*) A group of 12 students are found to have the following I.Q.'s:

112, 109, 125, 113, 116, 131, 112, 123, 108, 113, 132, 128

Is it reasonable to assume that these students have come from a large population whose mean I.Q.'s is 115?

Another group of 10 students are found to have the following I.Q.'s:

117, 110, 106, 109, 116, 119, 107, 106, 105, 108

Can we conclude that both the groups of students have come from the same population?

{ (*i*) Since $t_{11;0.025} = -2.201 < t = 1.386 < 2.201 = t_{11;0.975}$, we do not reject $H_0: \mu = 115$ against $H_1: \mu \neq 115$ at $\alpha = 0.05$.

(*ii*) Since $t = 2.608 > 2.086 = t_{20;0.975}$, we reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$; at $\alpha = 0.05$. }

(*c*) A random sample of 16 values from a normal population gave a mean of 42 inches and a sum of squared deviations from this mean as 135 (inches)^2. Test the hypothesis that the mean in the population is 43.5 inches.

Another random sample of 9 values from another normal population gave a mean of 41.5 inches and a sum of squares of deviations from this mean as 128 (inches)^2. Test the hypothesis that mean of first population equals the mean of the second population. assuming that the variances of the two populations are equal.

{ Since $t_{15;\,0.025} = -2.131 < t = -2 < 2.131 = t_{15;\,0.975}$, we do not reject $H_0: \mu = 43.5$ against $H_1: \mu \neq 43.5$.)

Since $t_{23;\,0.025} = -2.069 < t = 0.355 < 2.069 = t_{23;\,0.975}$, we do not reject $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ }

◆

13.5 INFERENCES ABOUT THE DIFFERENCE BETWEEN TWO POPULATION MEANS—*DEPENDENT SAMPLES*

There are many situations that require matched-pair comparisons. This method is appropriate when the observations of two dependent normal populations are compared. For example, if we are making inferences about the difference between blood pressure before and after administering a drug to heart patients, the blood pressure data after the drug is taken is dependent of the blood pressure before the drug is taken.

When matched samples are employed for making inference about the difference between two population means $\mu_2 - \mu_1$, it turns out that the procedures simplify to those for a single population mean μ. Suppose that we have an extreme and simplified form paired observations with pairing of before and after measurements. To analyse paired experiments, we consider for each pair the measurement before and after any conditions as variables X and Y, respectively.

Suppose that X is a normal random variable with mean μ_1 and variance σ_1^2, *i. e.*, $X \sim N(\mu_1, \sigma_1^2)$; Y is a normal random variable with mean μ_2 and variance σ_2^2, *i. e.*, $Y \sim N(\mu_2, \sigma_2^2)$; X and Y are dependent. We wish to estimate $\mu_2 - \mu_1$.

We consider for each pair the difference between the random variables X and Y and denote this random variable by D,

$$D = Y - X$$

We now consider the population of differences so obtained. We denote the mean of this population of differences by μ_D and variance by σ_D^2, which are given by

$$\mu_D = \mu_2 - \mu_1$$

$$\sigma_D^2 = \sigma_2^2 + \sigma_1^2 - 2\rho\,\sigma_1\,\sigma_2$$

A simple random sample of size n is selected from this population of differences. Let X_i and Y_i denote the before and after measurements respectively, for the *i-th* object in the random sample, so (X_i, Y_i) is a matched pair of observations.

Thus (X_1, Y_1), (X_2, Y_2), $\cdots$, (X_n, Y_n), is a random sample of n paired observations from the bivariate normal distribution with parameters given by $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = Var(X)$, $\sigma_2^2 = Var(Y)$, and $\rho = Corr(X, Y) = Cov(X, Y)/(\sigma_1\sigma_2)$.

The object is to make inferences about $\mu_D = \mu_2 - \mu_1$. To measure the change in measurements of the *i-th* object, we use the difference

$$D_i = Y_i - X_i, \qquad i = 1, 2, \cdots, n$$

then $D_1, D_2, \cdots, D_n$ are independently and identically distributed random variables with common normal distribution having mean μ_D and variance σ_D^2.

Now these differences $D_i = Y_i - X_i$, $i = 1, 2, \cdots, n$ may be thought of as a random sample of differences from a population of differences. We calculate the mean of the sample differences $D_1, D_2, \cdots, D_n$, denoted by $\overline{D}$. This statistic $\overline{D}$ is a random variable that has a sampling distribution with μ_D by the earlier procedures for a single population mean, but we are really estimating $\mu_D = \mu_2 - \mu_1$, the difference between the means of the two populations. The random variable

$$T = \frac{\overline{D} - \mu_D}{\hat{S}_D/\sqrt{n}}$$

follows a t-distribution with $\nu = n - 1$ degrees of freedom, where

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$\hat{S}_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1} = \frac{\sum_{i=1}^{n} D_i^2 - n\overline{D}^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} D_i^2 - \left(\sum_{i=1}^{n} D_i\right)^2 \Big/ n}{n-1}$$

$$= \frac{n\sum_{i=1}^{n} D_i^2 - \left(\sum_{i=1}^{n} D_i\right)^2}{n(n-1)}$$

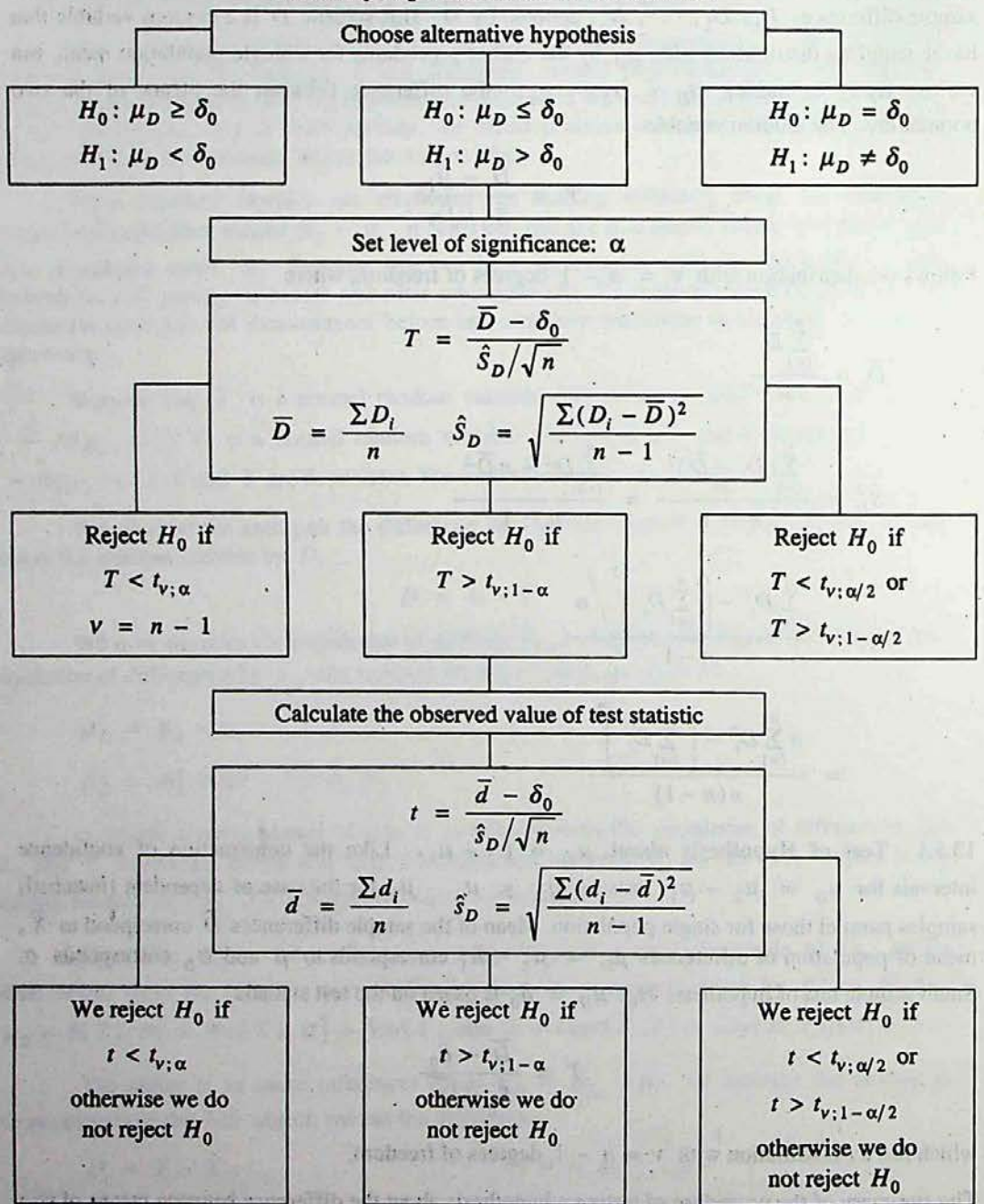
13.5.1 Test of Hypothesis about $\mu_D = \mu_2 - \mu_1$. Like the construction of confidence intervals for $\mu_D = \mu_2 - \mu_1$, tests on $\mu_D = \mu_2 - \mu_1$ for the case of dependent (matched) samples parallel those for single population. Mean of the sample differences $\overline{D}$ correspond to $\overline{X}$, mean of population of differences $\mu_D = \mu_2 - \mu_1$ corresponds to μ and σ_D corresponds σ. Small sample test of hypothesis $H_0: \mu_D = \delta_0$ is based on the test statistic

$$T = \frac{\overline{D} - \delta_0}{\hat{S}_D/\sqrt{n}}$$

which has a t-distribution with $\nu = n - 1$ degrees of freedom.

The summary of the procedure of testing a hypothesis about the difference between means of two normal populations with matched pairs of observations, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing difference between means of two normal populations, Dependent samples: Summary of procedure for 3 alternatives



Assumptions. To make valid inferences about the difference between means of two normal populations with matched pair of observations (dependent samples) the following assumptions must be met.

(*i*) A random sample of n differences must be selected from the population of differences

(*ii*) The population of differences must be (approximately) normally distributed.

The assumptions of an underlying normal distribution can be relaxed when the sample size is large. Applying the Central Limit Theorem to the differences $D_1, D_2, \cdots, D_n$ suggests a nearly normal distribution of $(\overline{D} - \mu_D)/(\hat{S}_D/\sqrt{n})$ when n is large (say greater than 30).

***Example* 13.24** *A new weight reducing technique, consisting of a liquid protein diet, is currently under going tests by the Food and Drug Administration (FDA) before its introduction into the market. A typical test performed by the FDA is the following. The weights of a random sample of 5 people are recorded before they are introduced to the liquid protein diet. The five individuals are then introduced to follow the liquid protein diet for 4 weeks. At the end of this period, their weights (in pounds) are again recorded. The results are listed in the table.*

Person	1	2	3	4	5
Weight before	150	195	188	197	204
Weight after	143	190	185	191	200

Perform a test of hypothesis at the 5% level of significance if the mean weight is smaller after the diet is used than before the diet is used.

Solution. The mean and standard deviation of the sample differences are

Person	Weight		Difference (after minus before)	
	Before	After		
i	x_i	y_i	$d_i = y_i - x_i$	d_i^2
1	150	143	−7	49
2	195	190	−5	25
3	188	185	−3	9
4	197	191	−6	36
5	204	200	−4	16
Sums			−25	135

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{-25}{5} = -5$$

$$\hat{s}_D = \sqrt{\frac{\Sigma d_i^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{135 - 5(-5)^2}{5-1}} = 1.58$$

The experimenter believes that the mean weight after is smaller than before, then the mean "after minus before" difference μ_D is less than zero. Hence $\mu_D < 0$ is the alternative hypothesis. The elements of the one-sided left tail test of hypothesis are:

Null hypothesis $H_0: \mu_D \geq 0$

Alternative hypothesis $H_1: \mu_D < 0$

Level of significance: $\alpha = 0.05$

Test statistic: $T = \dfrac{\overline{D} - \delta_0}{\hat{S}_D/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 5 - 1 = 4$

Critical value: $t_{\nu;\alpha} = t_{4;0.05} = -2.132$ (From Table 12)

Critical region: $T < -2.132$

Decision rule: Reject H_0 if $T < -2.132$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\overline{d} - \delta_0}{\hat{s}_D/\sqrt{n}} = \dfrac{-5-0}{1.58/\sqrt{5}} = -7.076$

Conclusion: Since $t = -7.076 < -2.132$, so we reject H_0.

***Example* 13.25** *Productivity (units produced per day) for a random sample of* 10 *workers was recorded before and after training. The following paired observations were obtained.*

Worker	1	2	3	4	5	6	7	8	9	10
Productivity before	54	56	50	52	55	52	56	53	53	60
Productivity after	60	59	57	56	56	58	62	55	54	64

Perform a test of hypothesis at the 1 *percent level of significance to determine if mean productivity is greater after training than before training.*

Solution. The mean and standard deviation of the sample differences are

Worker	Units produced per day		Difference (after minus before)		
	Before	After			
i	x_i	y_i	$d_i = y_i - x_i$	$d_i - \overline{d}$	$(d_i - \overline{d})^2$
1	54	60	6	2	4
2	56	59	3	−1	1
3	50	57	7	3	9
4	52	56	4	0	0
5	55	56	1	−3	9
6	52	58	6	2	4
7	56	62	6	2	4
8	53	55	2	−2	4
9	53	54	1	−3	9
10	60	64	4	0	0
Sums			$\sum d_i = 40$		$\sum (d_i - \overline{d})^2 = 44$

$$\overline{d} = \frac{\sum d_i}{n} = \frac{40}{10} = 4$$

$$\hat{s}_D = \sqrt{\frac{\sum (d_i - \overline{d})^2}{n-1}} = \sqrt{\frac{44}{10-1}} = 2.21$$

The experimenter believes that the mean productivity after is greater than before, then the population mean "after minus before" difference μ_D is greater than zero. Hence $\mu_D > 0$ is the alternative hypothesis. The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu_D \leq 0$

Alternative hypothesis $H_1: \mu_D > 0$

Level of significance: $\alpha = 0.01 \quad \Rightarrow \quad 1 - \alpha = 0.99$

Test statistic: $T = \dfrac{\overline{D} - \delta_0}{\hat{S}_D/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $v = n - 1 = 10 - 1 = 9$

Critical value: $t_{v;1-\alpha} = t_{9;0.99} = 2.821$ (From Table 12)

Critical region: $T > 2.821$

Decision rule: Reject H_0 if $T > 2.821$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\overline{d} - \delta_0}{\hat{s}_D/\sqrt{n}} = \dfrac{4 - 0}{2.21/\sqrt{10}} = 5.72$

Conclusion: Since $t = 5.72 > 2.821$, we reject H_0

***Example* 13.26** *A company is interested in hiring a new secretary. Several candidates are interviewed and the choice is narrowed to two possibilities. The final choice will be based on typing ability. Six letters are randomly selected from the company's file, and each candidate is required to type each one. The number of words typed per minute is recorded to each candidate. The data are listed in the following table.*

Letter	1	2	3	4	5	6
Candidate A	62	60	65	58	59	64
Candidate B	59	60	61	57	55	60

Do the date provide sufficient evidence to indicate a difference in the mean number of words typed per minute by the two candidates. Test using $\alpha = 0.02$.

Solution. The mean and standard deviation of the sample differences are

Letter	Number of words typed by		Difference (A minus B)	
	Candidate A	Candidate B		
i	x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	62	59	3	9
2	60	60	0	0
3	65	61	4	16
4	58	57	1	1
5	59	55	4	16
6	64	60	4	16
Sums			16	58

$$\bar{d} = \frac{\sum d_i}{n} = \frac{16}{6} = 2.667$$

$$\hat{s}_D = \sqrt{\frac{n\sum d_i^2 - (\sum d_i)^2}{n(n-1)}} = \sqrt{\frac{6(58) - (16)^2}{6(6-1)}} = 1.751$$

The experimenter believes that the mean typing rate differs for the candidate A and B, the population mean "A minus B" difference μ_D is not equal to zero. Hence $\mu_D \neq 0$ is the alternative hypothesis. The elements of the two-sided test of hypothesis are:

Null hypothesis $H_0: \mu_D = 0$

Alternative hypothesis $H_1: \mu_D \neq 0$

Level of significance: $\alpha = 0.02 \Rightarrow \alpha/2 = 0.01 \Rightarrow 1 - \alpha/2 = 0.99$

Test statistic: $T = \dfrac{\bar{D} - \delta_0}{\hat{S}_D/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 6 - 1 = 5$

Critical values: $t_{\nu;\alpha/2} = t_{5;0.01} = -3.365$,

$t_{\nu;1-\alpha/2} = t_{5;0.99} = 3.365$ (From Table 12)

Critical region: $T < -3.365$ or $T > 3.365$

Decision rule: Reject H_0 if $T < -3.365$ or $T > 3.365$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\bar{d} - \delta_0}{\hat{s}_D/\sqrt{n}} = \dfrac{2.667 - 0}{1.751/\sqrt{6}} = 3.731$

Conclusion: Since $t = 3.731 > 3.365$, so we reject H_0.

***Example* 13.27** *An experiment was performed with seven hop plants. One half of each plant was pollinated and the other half was not pollinated. The yield of the seed of each hop plant is tabulated as follows:*

Plant	*Pollinated*	*Non-pollinated*
1	0.78	0.21
2	0.76	0.12
3	0.43	0.32
4	0.92	0.29
5	0.86	0.30
6	0.59	0.20
7	0.68	0.14

Determine at the 5% level whether the pollinated half of the plant gives a higher yield in seed than the non-pollinated half. State the assumptions and hypothesis to be tested and carry through the computations to make a decision

Solution. The mean and standard deviation of the sample differences are

Plant	Yield in seed		Difference (Pollinated minus Non-pollinated)	
	Pollinated	Non-pollinated		
i	x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	0.78	0.21	0.57	0.3249
2	0.76	0.12	0.64	0.4096
3	0.43	0.32	0.11	0.0121
4	0.92	0.29	0.63	0.3969
5	0.86	0.30	0.56	0.3136
6	0.59	0.20	0.39	0.1521
7	0.68	0.14	0.54	0.2916
Sums			3.44	1.9008

$$\bar{d} = \frac{\sum d_i}{n} = \frac{3.44}{7} = 0.491$$

$$\hat{s}_D = \sqrt{\frac{n\sum d_i^2 - (\sum d_i)^2}{n(n-1)}} = \sqrt{\frac{7(1.9008) - (3.44)^2}{7(7-1)}} = 0.1872$$

The experimenter believes that the pollinated half gives a higher mean yield in seed, then the alternative hypothesis is $\mu_D > 0$. The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \mu_D \leq 0$

Alternative hypothesis $H_1: \mu_D > 0$

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $T = \dfrac{\bar{D} - \delta_0}{\hat{S}_D/\sqrt{n}}$ follows a t-distribution under H_0 with

Degrees of freedom: $\nu = n - 1 = 7 - 1 = 6$

Critical values: $t_{\nu;1-\alpha} = t_{6;0.95} = 1.943$ (From Table 12)

Critical region: $T > 1.943$

Decision rule: Reject H_0 if $T > 1.943$, otherwise do not reject H_0.

Observed value: $t = \dfrac{\bar{d} - \delta_0}{\hat{s}_D/\sqrt{n}} = \dfrac{0.491 - 0}{0.1872/\sqrt{7}} = 6.939$

Conclusion: Since $t = 6.939 > 1.943$, so we reject H_0, and conclude that the data provides a sufficient evidence that pollination gives a higher mean yield in seed.

Exercise 13.5

1. (*a*) Haemoglobin values were determined on six patients before starting and after three weeks on B_{12} Therapy. The following data were obtained

Individual Number	Haemoglobin (gm) Before Therapy	Haemoglobin (gm) After Therapy
1	12.2	13.0
2	11.3	13.4
3	14.7	16.0
4	11.4	13.6
5	11.5	14.0
6	12.7	13.8

Do the data indicate a significant improvement ?

(Since $t = 5.927 > 2.015 = t_{5;0.95}$, we reject $H_0: \mu_D \leq 0$ in favour of $H_1: \mu_D > 0$)

(*b*) Eleven school boys were given a test in Drawing. They were given one month's further tuition and a second test of equal difficulty was held at the end of it.

Marks in 1st test	23	20	19	21	18	20	18	17	23	16	19
Marks in 2nd test	24	19	22	18	20	22	20	20	20	20	17

Do the marks give evidence that the students have benefited by the extra coaching?

(Since $t = 0.956 < 1.812 = t_{10;0.95}$, so we do not reject $H_0: \mu_D \leq 0$ against $H_1: \mu_D > 0$)

2. (*a*) A taxi company is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption in km per litre, was recorded as follows:

Radial tires	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
Belted tires	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9

At the 0.025 level of significance, can we conclude, that cars equipped with radial tires give better fuel economy than those equipped with belted tires ? Assume the population to be normally distributed.

(Since $t = 2.490 > 2.201 = t_{11;0.975}$, we reject $H_0: \mu_D \leq 0$ in favour of $H_1: \mu_D > 0$)

(*b*) A certain stimulus administered to each of the nine patients resulted in the following increase in blood pressure:

5, 1, 8, 0, 3, 3, 5, −2, 4

Can it be concluded that blood pressure in general is increased by ad stimulus?

(Since $t = 3.0 > 1.86 = t_{8;\,0.95}$, we reject $H_0: \mu_D \leq 0$ in favour of

3. **(*a*)** To verify whether a course in Statistics improved performance a similar to 12 participants both before and after the course. The original grade alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 4 and 72. After the course the grades were, in the same order, 53, 38, 69 73, 48, 73, 74, 60 and 78.

(*i*) Was the course useful, as measured by performance on the test ? C 12 participants as a sample from a population.

(*ii*) Would the same conclusion be reached if the tests were not consid (use 5% level of significance in both cases).

{ (*i*) Since $t = 3.445 > 1.796 = t_{11;\,0.95}$, we reject $H_0: \mu_D \leq 0$ $H_1: \mu_D > 0$.

(*ii*) Since $t = 0.863 < 1.717 = t_{22;\,0.95}$, we do not reject $H_0: \mu_2$ against $H_1: \mu_2 - \mu_1 > 0$ }.

(*b*) The time required by 10 persons to perform a task in seconds before and afte a mild stimulant are given in the accompanying table.

Before	34	45	31	43	40	41	33	29	41
After	29	42	32	29	36	42	26	28	38

Test the hypothesis that there is no difference between the mean times in the and "after" populations. As an alternative, assumed that the after population w lower mean. Use 5% level of significance.

(Since $t = -2.83 < -1.833 = t_{9;\,0.05}$, we reject $H_0: \mu_D \geq 0$ in $H_1: \mu_D < 0$.)

4. **(*a*)** Ten young recruits were put through a physical training programme by the arm weights were recorded before and after the training with the following results:

Recruit	1	2	3	4	5	6	7	8	9
Weight (Before)	127	126	162	170	143	205	168	175	197
Weight (After)	135	200	160	182	147	200	172	186	193

Using $\alpha = 0.05$, should we conclude that the training programme affects the a weight of young recruits.

(Since $t_{9;\,0.025} = -2.262 < t = 1.471 < 2.262 = t_{9;\,0.975}$, we do not $H_0: \mu_D = 0$ against $H_1: \mu_D \neq 0$.)

(*b*) The following data give paired yields of two varieties of wheat. Each pair was plan a different locality.

Variety I	45	32	58	57	60	38	47	51	42
Variety II	47	34	60	59	63	44	49	53	46

Test the hypothesis that the mean yields are equal.

(Since $t = 6.725 > 2.262 = t_{9;\,0.975}$, we reject H_0: $\mu_D = 0$ in favour of H_1: $\mu_D \neq 0$.)

5. (*a*) Twenty college freshmen were divided into 10 pairs, each member of the pair having approximately the same I.Q. One of each pair was selected at random and assigned to a Mathematics section using programmed materials only. The other members of each pair were assigned to a section in which the teacher lectured. At the end of the semester each group was given the same examination and the following results were recorded:

Pair	Programmed materials	Lectures
1	76	81
2	70	52
3	85	87
4	58	70
5	91	86
6	75	77
7	82	90
8	64	73
9	79	85
10	88	83

Test the hypothesis that there is no difference between the mean scores in the "programmed material" and "lectures" populations. Use 5% level of significance.

(Since $t_{9;\,0.025} = -2.262 < t = 0.571 < 2.262 = t_{9;\,0.975}$, we do not reject H_0: $\mu_D = 0$ against H_1: $\mu_D \neq 0$.)

(*b*) It is claimed that a new diet will reduce a person's weight by at least 10 pounds on the average in a period of 2 weeks. The weight of 7 women who followed this diet were recorded before and after a 2-week period.

Woman	1	2	3	4	5	6	7
Weight (Before)	129	133	136	152	141	138	125
Weight (After)	130	121	128	137	129	132	120

Test the manufacturer's claim at a 5% level of significance for the mean difference in weights. Assume the distribution of weights before and after to be approximately normal.

(Since $t = 0.910 < 1.943 = t_{6;\,0.95}$, we do not reject H_0: $\mu_D \leq -10$ against H_1: $\mu_D > -10$)

◆

13.6 TEST OF HYPOTHESIS ABOUT THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS, $\pi_1 - \pi_2$

There are many economic and management problems where we must decide whether observed differences between two sample proportions come from common or different populations. Such problems require a comparison between the rates of incidence of a characteristic in two populations.

13.6.1 Forms of Hypothesis. We are interested in tests about the parameter $\pi_1 - \pi_2$. Let δ_0 be the hypothesized value of the difference between two population proportions, then the three possible null hypotheses about the difference between population proportions, and their corresponding alternative hypotheses, are:

1. $H_0: \pi_1 - \pi_2 \geq \delta_0$ against $H_1: \pi_1 - \pi_2 < \delta_0$
2. $H_0: \pi_1 - \pi_2 \leq \delta_0$ against $H_1: \pi_1 - \pi_2 > \delta_0$
3. $H_0: \pi_1 - \pi_2 = \delta_0$ against $H_1: \pi_1 - \pi_2 \neq \delta_0$

Depending upon the alternative hypothesis a one-sided left tail test is required for (1); (2) requires a one-sided right tail test, and (3) requires a two-tail test.

13.6.2 Test based on Normal Distributions. The unknown proportion of elements possessing the particular characteristic in population I and in population II are denoted by π_1 and π_2, respectively. A random sample of size n_1 is taken from population I and the number of successes is denoted by X_1. An independent random sample of size n_2 is taken from population II and the number of successes is denoted by X_2. The sample proportions are:

$$P_1 = \frac{X_1}{n_1} \qquad \text{and} \qquad P_2 = \frac{X_2}{n_2}$$

An intuitively appealing estimator for $\pi_1 - \pi_2$ is the difference between the sample proportions $P_1 - P_2$. When testing hypothesis about $\pi_1 - \pi_2$, we will use the sampling distribution of $P_1 - P_2$. The sampling distribution of $P_1 - P_2$ will have mean and standard error as

$$\mu_{P_1 - P_2} = \pi_1 - \pi_2, \qquad \sigma_{P_1 - P_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

For large sample sizes n_1 and n_2, the random variable

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

is approximately standard normal. The estimate of the standard error of $P_1 - P_2$ can be obtained by replacing π_1 and π_2 by their sample estimates P_1 and P_2 as

$$\hat{\sigma}_{P_1 - P_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

The random variable Z then becomes

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$$

(a) ***Testing the Hypothesis that the Difference between Two Population Proportions Equals Some Non-zero Value.*** To test the more general hypothesis $H_0: \pi_1 - \pi_2 = \delta_0$, we use the test statistic

$$Z = \frac{(P_1 - P_2) - \delta_0}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$$

(b) ***Testing the Hypothesis that Two Independent Populations have Same Proportion of Successes.*** Our aim is to test the null hypothesis of no difference $H_0: \pi_1 = \pi_2$. Under the null hypothesis $H_0: \pi_1 - \pi_2 = 0$, the two populations have equal proportions $\pi_1 = \pi_2$, we denote the unspecified common population proportion by π.

Since under the null hypothesis it is assumed that $\pi_1 = \pi_2 = \pi$, then the sampling distribution of $P_1 - P_2$ will have mean

$$\mu_{P_1 - P_2} = \pi - \pi = 0$$

and standard error

$$\sigma_{P_1 - P_2} = \sqrt{\frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2}} = \sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The unknown population proportion π involved in the standard error must now be replaced by the sample proportion.

The difficulty in making this replacement lies in the fact that we have two estimates of π, P_1 and P_2, since two different samples were collected. Which of these two estimates should be used to estimate the unknown population proportion π.

Since we wish to obtain the best estimate available, it would seem reasonable to use an estimator that would pool the information from the two samples.

The proportion of successes in the combined sample provides the pooled estimate.

$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

The estimate of standard error then becomes

$$\hat{\sigma}_{P_1 - P_2} = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

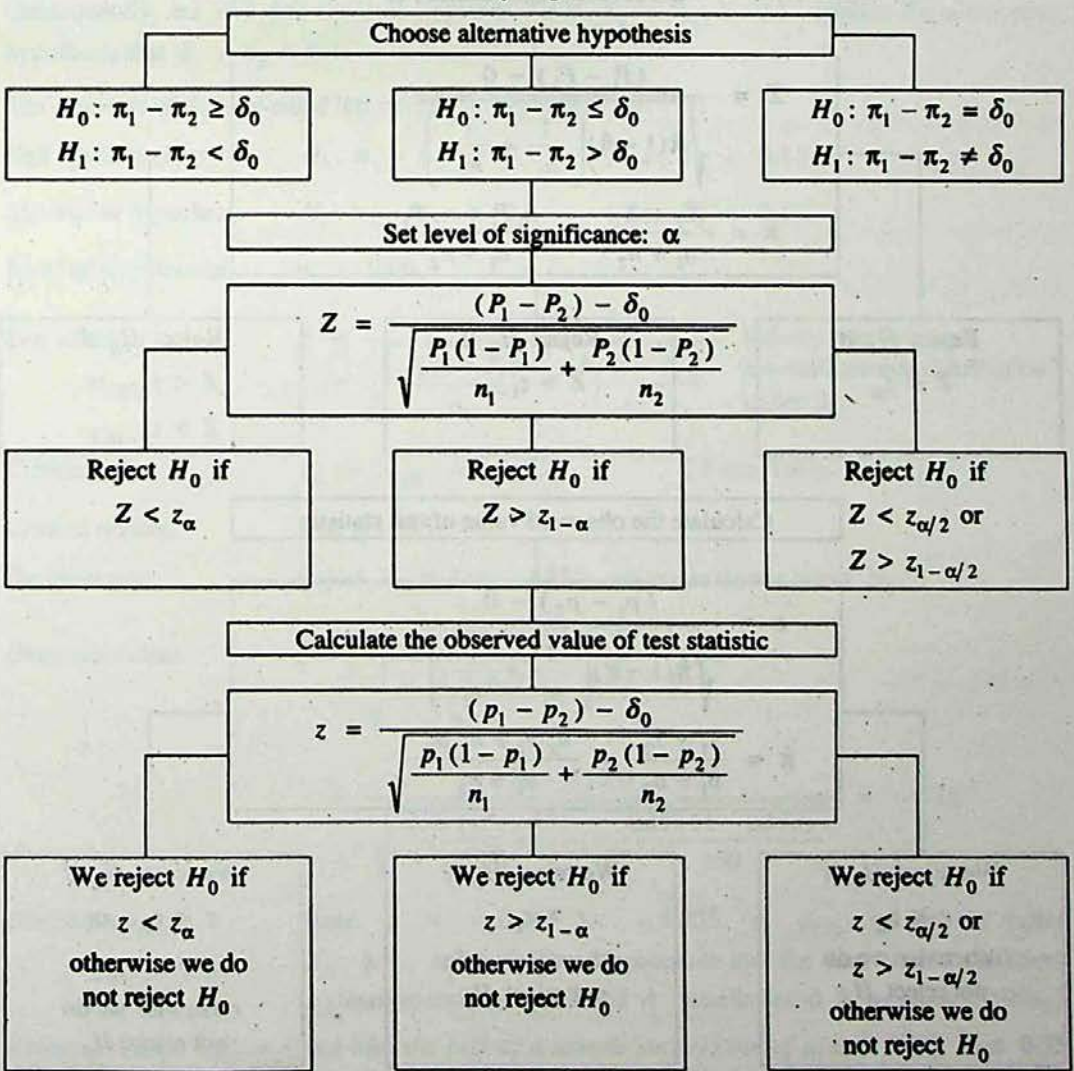
For large sample sizes n_1 and n_2, the random variable

$$Z = \frac{P_1 - P_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is approximately standard normal under the null hypothesis $H_0: \pi_1 - \pi_2 = 0$.

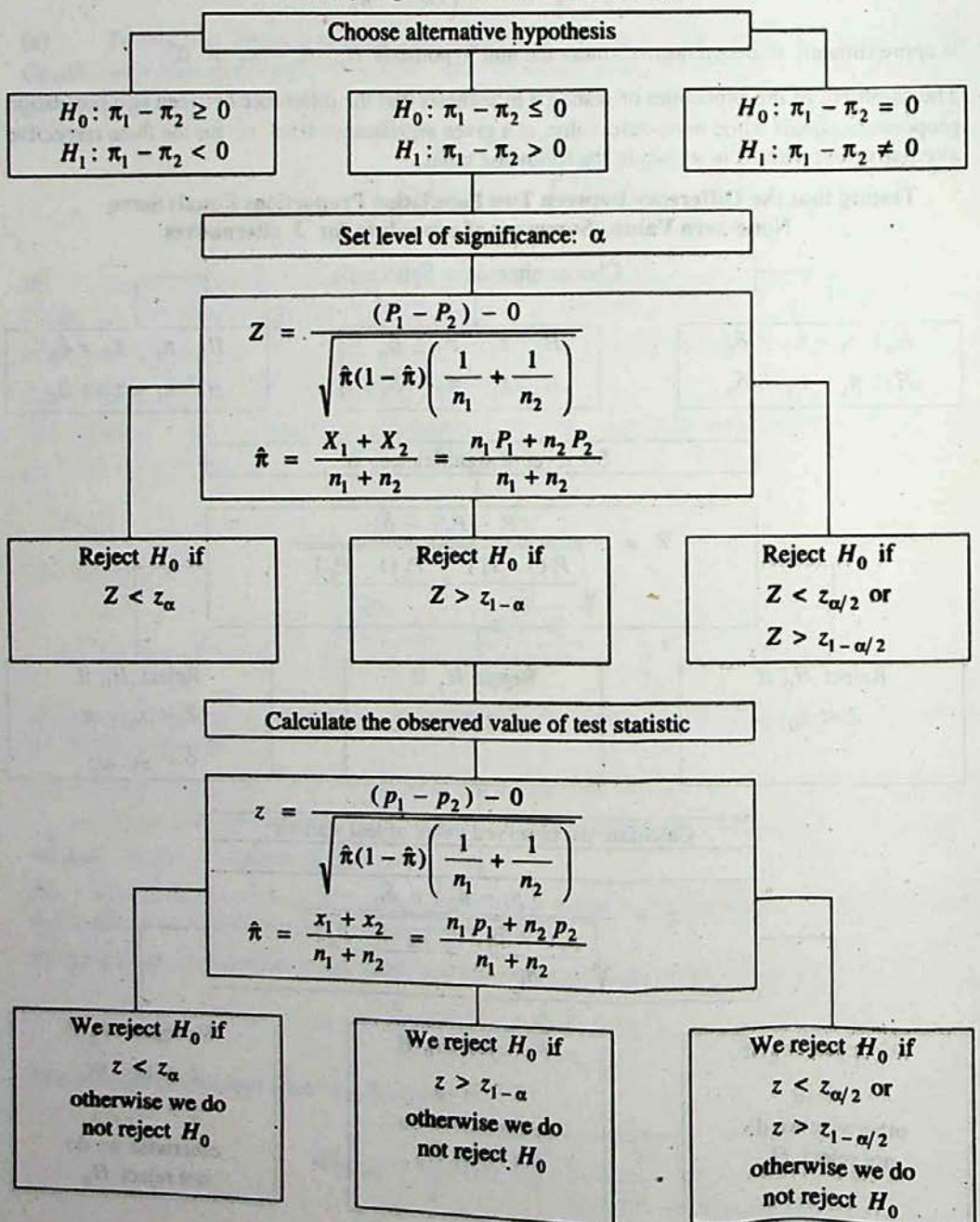
The summary of the procedure of testing a hypothesis that the difference between two population proportions equals some none-zero value, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing that the Difference between Two Population Proportions Equals Some None-zero Value. Summary of procedure for 3 alternatives



The summary of the procedure of testing a hypothesis that two independent populations have same proportion of successes, at a given significance level α, for the three respective alternative hypotheses is shown in the following table.

Testing that Two Independent Populations have Same Proportion of Successes: Summary of procedure for 3 alternatives



***Example* 13.28** *A cigarette manufacturing firm distributes two brands of cigarettes. It is found that* 56 *of* 200 *smokers prefer brand "A" and that* 29 *of* 150 *smokers prefer brand "B". Test at* 0.06 *level of significance that brand "A" outsell brand "B" by at least* 10% *against the alternative hypothesis that the difference is less than* 10%.

Solution. We have $n_1 = 200, \quad x_1 = 56, \quad p_1 = \frac{x_1}{n_1} = \frac{56}{200} = 0.28$

$n_2 = 150, \quad x_2 = 29, \quad p_2 = \frac{x_2}{n_2} = \frac{29}{150} = 0.193$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the difference between the two proportions of smokers $\pi_1 - \pi_2$ is less than 0.1. Consequently, we will test the null hypothesis that $\pi_1 - \pi_2 \geq 0.1$ against the alternative hypothesis that $\pi_1 - \pi_2 < 0.1$.

The elements of the one-sided left tail test of hypothesis are:

Null hypothesis $H_0: \pi_1 - \pi_2 \geq 0.1 \quad (i.e., \delta_0 \geq 0.1)$

Alternative hypothesis $H_1: \pi_1 - \pi_2 < 0.1 \quad (i.e., \delta_0 < 0.1)$

Level of significance: $\alpha = 0.06$

Test statistic: $Z = \dfrac{(P_1 - P_2) - \delta_0}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_2)}{n_2}}}$ follows approximately standard normal distribution under H_0

Critical value: $z_\alpha = z_{0.06} = -1.555$ { From Table 10 (*a*) }

Critical region: $Z < -1.555$

Decision rule: Reject H_0 if $Z < -1.555$, otherwise do not reject H_0.

Observed value:

$$z = \frac{(p_1 - p_2) - \delta_0}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

$$= \frac{(0.28 - 0.193) - 0.10}{\sqrt{\dfrac{0.28\,(1-0.28)}{200} + \dfrac{0.193\,(1-0.193)}{150}}} = -0.287$$

Conclusion: Since $z = -0.287 > -1.555 = z_{0.06}$, we do not reject $H_0: \pi_1 - \pi_2 \geq 0.10$ and conclude that the data present sufficient evidence to indicate that brand *A* outsells brand *B* by at least 10%.

***Example* 13.29** *A firm found with the help of a sample survey (size of sample* 900) *that* 0.75 *of the population consumes things produced by them. The firm then advertised the goods in paper*

and on radio. After one year, sample of size 1000 *reveals that proportion of consumers of the goods produced by the firm is now 0.8. Is this rise significant indicating that the advertisement was effective?*

Solution. We have $n_1 = 900, \quad p_1 = 0.75, \quad n_2 = 1000, \quad p_2 = 0.80$

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$= \frac{900(0.75) + 1000(0.80)}{900 + 1000} = 0.776$$

The objective of the sampling is to attempt to support the research (alternative) hypothesis that the proportion of consumers after advertisement π_2 is greater than that of before π_1. Consequently, we will test the null hypothesis that $\pi_2 \leq \pi_1$ against the alternative hypothesis that $\pi_2 > \pi_1$.

The elements of the one-sided right tail test of hypothesis are:

Null hypothesis $H_0: \pi_2 \leq \pi_1 \Rightarrow \pi_2 - \pi_1 \leq 0 \; (i.e., \delta_0 \leq 0)$

Alternative hypothesis $H_1: \pi_2 > \pi_1 \Rightarrow \pi_2 - \pi_1 > 0 \; (i.e., \delta_0 > 0)$

Level of significance: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Test statistic:
$$Z = \frac{(P_2 - P_1) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
follows approximately standard normal distribution under H_0

Critical value: $z_{1-\alpha} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

Critical region: $Z > 1.645$

Decision rule: Reject H_0 if $Z > 1.645$, otherwise do not reject H_0.

Observed value:
$$z = \frac{(p_2 - p_1) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{(0.8 - 0.75) - 0}{\sqrt{0.776(1-0.776)\left(\frac{1}{900} + \frac{1}{1000}\right)}} = 2.61$$

Conclusion: Since $z = 2.61 > 1.645 = z_{0.95}$, we reject H_0 and conclude that the data present sufficient evidence to indicate the consumption after advertisement is higher than that of before.

Exercise 13.6

1. (*a*) For each of the following sets of data, test the hypothesis that there is a common proportion π.

	Sample I		Sample II		Hypotheses	Level of significance
	Sample size	Number of Successes	Sample size	Number of Successes		
(*i*)	150	125	200	176	$H_0: \pi_1 \geq \pi_2$ $H_1: \pi_1 < \pi_2$	5%
(*ii*)	1000	542	900	427	$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$	1%
(*iii*)	100	63	120	80	$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$	10%
(*iv*)	80	40	60	23	$H_0: \pi_1 \leq \pi_2$ $H_1: \pi_1 > \pi_2$	5%

{ (*i*) Since $z = -1.245 > -1.645 = z_{0.05}$, we do not reject $H_0: \pi_1 \geq \pi_2$ against $H_1: \pi_1 < \pi_2$.

(*ii*) Since $z = 2.941 > 2.576 = z_{0.995}$, we reject $H_0: \pi_1 = \pi_2$ in favour of $H_1: \pi_1 \neq \pi_2$.

(*iii*) Since $z_{0.05} = -1.645 < z = -0.573 < 1.645 = z_{0.95}$, we do not reject $H_0: \pi_1 = \pi_2$ against $H_1: \pi_1 \neq \pi_2$.

(*iv*) Since $z = 1.373 < 1.645 = z_{0.95}$, we do not reject $H_0: \pi_1 \leq \pi_2$ against $H_1: \pi_1 > \pi_2$. }

(*b*) In a population that have certain minor blood disorder, samples of 100 males and 100 females are taken. It is found that 31 males and 24 females have the blood disorder. Can we conclude at 0.01 level of significance that proportion of men who have blood disorder is greater than proportion of women?

{ Since $z = 1.109 < 2.326 = z_{0.99}$, we do not reject $H_0: \pi_1 \leq \pi_2$ against $H_1: \pi_1 > \pi_2$. }

2. (*a*) The records of a hospital show that 52 men in a sample of 1000 men versus 23 women in a sample of 1000 women were admitted because of heart disease. Do these data present sufficient evidence to indicate a higher rate of heart disease among men admitted to the hospital?

{ Since $z = 3.413 > 1.645 = z_{0.95}$, we reject $H_0: \pi_1 \leq \pi_2$ in favour of $H_1: \pi_1 > \pi_2$. }

(*b*) In a study to estimate the proportion of housewives who own an automatic dryer, it is found that 63 of 100 urban residents have a dryer and 59 of 125 suburban residents own a dryer. Is there a significant difference between the proportions of urban and suburban housewives who own an automatic dryer? (use a 0.04 level of significance).
{ Since $z = 2.364 > 2.0537 = z_{0.98}$, so we reject $H_0: \pi_1 = \pi_2$ in favour of $H_1: \pi_1 \neq \pi_2$. }

3. (*a*) A random sample of 150 light bulbs manufactured by a firm A showed 12 defective bulbs while a random sample of 100 light bulbs manufactured by another firm B showed 4 defective bulbs. Is there a significant difference between the proportions defectives of the two firms?
{ Since $z_{0.025} = -1.960 < z = 1.270 < 1.960 = z_{0.975}$, we do not reject $H_0: \pi_1 = \pi_2$ against $H_1: \pi_1 \neq \pi_2$. }

(*b*) In a random sample of 800 adults from the population of a large city 600 are found to be smokers. In a random sample, of 1000 adults from another large city, 700 are smokers. Do the data indicate that the cities are significantly different with respect to the prevalence of smoking among men?

{ Since $z = 2.353 > 1.960 = z_{0.975}$, we reject $H_0: \pi_1 = \pi_2$ in favour of $H_1: \pi_1 \neq \pi_2$. }

(*c*) In two large populations, there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in a sample of 1200 and 900 respectively from the two populations?
{ Since $z = 2.529 > 1.960 = z_{0.975}$, we reject $H_0: \pi_1 = \pi_2$ in favour of $H_1: \pi_1 \neq \pi_2$ at $\alpha = 0.05$. }

4. (*a*) A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect article in a batch of 100. Has the machine been improved? Use 5% level of significance?'
(Since $z = -0.104 > -1.645 = z_{0.05}$, we do not reject $H_0: \pi_2 \geq \pi_1$ against $H_1: \pi_2 < \pi_1$)

(*b*) In a random sample of 250 persons who skipped breakfast 102 reported that they experienced mid morning fatigue, and in a random sample of 250 persons who ate breakfast, 73 reported that they experienced mid morning fatigue. Use 0.01 level of significance to test the null hypothesis that there is no difference between the corresponding proportions against the alternative hypothesis that the mid morning fatigue is more prevalent among persons who skip breakfast.
(Since $z = 2.719 > 2.326 = z_{0.99}$, we reject $H_0: \pi_1 \leq \pi_2$ in favour of $H_1: \pi_1 > \pi_2$)

5. (*a*) A manufacturer of house-dresses sent out advertising by mail. He sent samples of material to each of two groups of 1000 women. For one group he enclosed a white return envelope and for the other group, a blue envelope. He received orders from 9% and 12% respectively. Is it quite certain that the blue envelope will help sales.

Use $\alpha = 0.05$.
(Since $z = 2.19 > 1.645 = z_{0.95}$, we reject $H_0: \pi_2 \leq \pi_1$ in favour of $H_1: \pi_2 > \pi_1$)

(b) A random sample of 150 high school students was asked whether they would turn to their father or their mother for help with a homework assignment in Mathematics and another random sample of 150 high school students was asked the same question with regard to a homework assignment in English. Use the result shown in the following table and the 0.01 level of significance to test whether or not there is a difference between the true proportions of high school students who turn to their fathers rather than their mothers for help in these two subjects:

	Mathematics	*English*
Mother:	59	85
Father:	91	65

(Since $z = 3.016 > 2.576 = z_{0.995}$, we reject $H_0: \pi_1 = \pi_2$ in favour of $H_1: \pi_1 \neq \pi_2$)

◆

Exercise 13.7

Objective Questions

1. Fill in the blanks.

(*i*) A statistical ———— is an assertion about the distribution of one or more random variables. (*hypothesis*)

(*ii*) A statistical hypothesis ———— is a procedure to determine whether or not an assumption about some parameter of population is supported by an observed random sample. (*testing*)

(*iii*) A ———— hypothesis is that hypothesis which is tested for possible rejection under the assumption that it is true. (*null*)

(*iv*) An ———— hypothesis is that hypothesis which we are willing to accept when the null hypothesis is rejected. (*alternative*)

(*v*) The ———— hypothesis always contains some form of an equality sign. (*null*)

(*vi*) The ———— hypothesis never contains the sign of equality and is always in an inequality from. (*alternative*)

(*vii*) A statistic on the basis of which a decision is made about the hypotheses of interest is called ————. (*test statistic*)

(*viii*) A ———— region specifies a set of values of the test statistic for which the H_0 is rejected. (*rejection*)

(*ix*) An ———— region specifies a set of values of the test statistic for which the H_0 is not rejected. (*accepting*)

(*x*) The values of the test statistic which separate the rejection region from acceptance region are called ———— values. (*critical*)

(*xi*) If the critical region is located equally in both tails of the sampling distribution of test statistic, the test is called ———— test. (*two-tailed*)

(*xii*) If the critical region is located in only one tail of the sampling distribution of test statistic, the test is called ———— test. (*one-tailed*)

2. Fill in the blanks.

(*i*) A value of the test statistic is said to be statistically ———— if it falls in the acceptance region. (*insignificant*)

(*ii*) A value of the test-statistic is said to be statistically significant if it falls in the ———— region. (*critical*)

(*iii*) If the null hypothesis is false, we may accept it leading to a ———— decision. (*wrong*)

(*iv*) If the null hypothesis is false, we may reject it leading to a ———— decision. (*correct*)

(*v*) A ———— error is made by rejecting H_0 if H_0 is actually true. (*Type*-I)

(*vi*) A ———— error is made by accepting H_0 if H_1 is actually true. (*Type*-II)

(*vii*) The level of ———— of a test is the maximum probability with which we are willing to a risk of Type-I error. (*significance*)

(*viii*) The level of ———— is the probability of accepting a true null hypothesis. (*confidence*)

(*ix*) The ———— of freedom is the number of independent or freely chosen variables. (*degrees*)

3. Mark off the following statements as True or False.

(*i*) The types of statistical inferences are estimation of parameters and testing of hypotheses. (*true*)

(*ii*) A null hypothesis is rejected when a test statistic has a value that is not consistent with the null hypothesis. (*true*)

(*iii*) The probability of accepting the true null hypothesis is called the level of significance. (*false*)

(*iv*) The probability of rejecting the true null hypothesis is called the level of confidence. (*false*)

(*v*) The probability of accepting the true null hypothesis is called the level of confidence. (*true*)

(*vi*) An assumption made about the population parameter which may or may not be true is called statistical hypothesis. (*true*)

(*vii*) The null hypothesis and alternative hypothesis are complementary to each other. (*true*)

(*viii*) The null hypothesis H_0 always contain some from of an equality sign. (*true*)

(*ix*) The alternative H_1 never contains the sign of equality. (*true*)

4. Mark off the following statements as True or False.

(*i*) The level of significance is the probability of accepting a null hypothesis when it is true. (*false*)

(*ii*) A null hypothesis is rejected if the value of test-statistic is consistent with the H_0. (*false*)

(*iii*) The Type-I error is considered more serious than a Type-II error. (*true*)

(*iv*) A value of the test-statistic is said to be statistically insignificant if it falls in the rejection region. (*false*)

(*v*) A value of the test statistic is said to be statistically significant if it falls in the rejection region. (*true*)

(*vi*) If the critical region is located in only one tail of the sampling distribution of the test-statistic, then it is called one tailed or one-sided test. (*true*)

(*vii*) If the critical region is equally located in both tails of the sampling distribution of the test-statistic, then it is called a two-tailed or two sided test. (*true*)

(*viii*) The degrees of freedom is the number of dependent variables. (*false*)

(*ix*) The *t*-distribution approaches the normal distribution as the sample size increases. (*true*)

(*x*) The standardized normal distribution has smaller dispersion than student's *t*-distribution. (*true*)

(*xi*) H_0 is rejected when probability of its occurrence is equal to or less than level of significance. (*true*)

14 SIMPLE LINEAR REGRESSION AND CORRELATION

14.1 RELATIONS BETWEEN VARIABLES

The concept of a relation between two variables such as family incomes and family expenditures for housing, is a familiar one. We now distinguish between a *functional relation* and a *statistical relation*, and consider each of them in turn.

14.1.1 Functional Relation between Two Variables. A *functional relation* between two variables, is a perfect relation, where the value of the dependent variable is uniquely determined from the value of the independent variable. A functional relation is expressed by a mathematical formula. If x is the *independent* variable and y is the *dependent* variable, a functional relation is of the form

$$y = f(x)$$

Given a particular value of x, the function $f(x)$ gives the corresponding value of y. The observations, when plotted on a graph, all fall directly on the line or curve of the functional relationship. This is the main characteristic of all functional relationships.

14.1.2 Statistical Relation between Two Variables. A *statistical relation* is a relation where the value of the dependent variable is not uniquely determined when the level of the independent variable is specified. A statistical relation, unlike a functional relation, is not exact. The value of y is not uniquely determined from knowledge of x. The observations, when plotted on a graph, do not fall directly on the line or curve of the relationship. This is the main characteristic of all statistical relationships.

In many fields such as business, economics and administration exact relations are not generally observed among the variables, but rather statistical relationships prevail. For example: (*i*) The grade point Y secured by a student in the college is undoubtedly related to his grade point x secured in the school. (*ii*) The consumption expenditure Y of a household is related to its income x. (*iii*) The maintenance cost Y per year for an automobile is related to his age x. (*iv*) The yield Y of wheat is related to the quantity x of a fertilizer. (*v*) The amount of sales Y of a newly produced item may be related to its advertising cost x. (*vi*) The weight Y of a baby is certainly related to his age x. (*vii*) The saving Y of a person or a firm is related to his/its income. (*viii*) The height Y of a son is undoubtedly related to the height x of his father, *etc.*

Causal Relation. Another factor to consider is whether a causal relationship exists between two variables. In the example of steel output and labour input, it is clear that a causal relationship does exist. The number of workers will influence the number of tons of steel produced. There is also a causal relationship between hours of sunshine and the rate of growth of tulips. Conversely, it is less clear that more steel output will cause a rise in the number of workers, nor will we make the sunshine more by forcing tulips to grow faster. It is important to note regression analysis and correlation analysis make no assertions about causality.

14.2 REGRESSION ANALYSIS

One of the most common and important tasks that statisticians must face is to determine the existence and nature of relationships between variables in a problem. We are interested in relationships between variables because we may often possess information about some variables and wish to use that information to draw conclusions about another variable. In many situations, we face the problems that involve two or more variables and we are to make inferences about how the changes in one variable are related to the changes in other variables, and how one set of variables is considered to predict or account for the other variable. These problems can be dealt with measuring statistical relationships between variables, representing the relationships in mathematical (functional) form and evaluating the significance of the relationships.

The *regression analysis* provides a method of estimating an average relationship (often linear) between two or more variables, which allows the investigator to explain and predict and this is, in a sense, the best possible approximation. The regression analysis provides an equation that can be used for estimating the average value of one variable from given values of other variables.

14.2.1 Simple Regression. The *simple regression* is a relationship that describes the dependence of the expected value of the dependent random variable for a given value of the independent non-random variable. In statistical relationships, if only two values are involved:

Regressor. The variable, that forms the basis of estimation or prediction, is called the regressor. It is also called as the predictor variable or independent variable or controlled variable or explanatory variable. It is usually denoted by x.

Regressand. The variable, whose resulting value depends upon the selected value of the independent variable, is called the regressand. It is also called as the response variable or the predictand variable or dependent variable or explained variable. It is usually denoted by Y.

The values of the independent variable x are determined by the experimenter and they are fixed in advance. They are arbitrarily selected constants and thus have no error attached with them. The independent variable is not random but a mathematical variable and we can choose the values we give to it. On the other hand, however, the problem is usually complicated by the fact that the dependent variable is subject to experimental variation or scatter. Besides depending upon the regressor variable, there is a random error in determining the response variable. Thus the response variable possesses a random character, it is left free to take on any value that may be possibly associated to a given value of the independent variable.

Let Y be the response variable and x be the regressor variable and $\mu_{Y|x} = E(Y|x)$ be the expected value of the distribution of the random variable Y for a given value of the non-random variable x, then the simple regression is given by

$$\mu_{Y|x} = f(x)$$

where $f(x)$ is a function that describes the relationship between the regressor x and the response Y and $f(x)$ may be of linear, quadratic, exponential, geometric, or any other form.

14.2.2 Regression Function. When we look for a relationship $\mu_{Y|x} = f(x)$, where the function $f(x)$ is to be determined, *i. e.*, given the points only we have to 'work backwards' or regress to the original function $f(x)$. Hence this function is called *regression function.*

14.2.3 Regression Curve. The *regression curve* is the locus (a continuous set of points) of the expected value of the response variable for given values of the regressor variable. If several measurements are made on the response variable Y at the same value of the regressor variable x, then the results will form a distribution. The curve which joins the expected values of these distributions for different values of x is called the simple regression curve of Y on x.

14.3 CURVE FITTING

Curve fitting is a process of estimating, from an observed sample, the parameters of the population regression function of a response variable on a regressor variable.

14.3.1 Least Squares Principle. The *principle of least squares* says that the sum of squares of the residuals of observed values from their corresponding estimated values should be the least possible. This principle was given by a French mathematician Adrien Legendre.

14.3.2 Least Squares Fit. Among all the curves approximating a given data, the curve is called a *least squares fit* for which the sum of squares of the residuals of the observed values from their corresponding estimated values is the least.

For a given set of observed data, different curves have different values of the sum of squares of the residuals. The best fitting curve is the one having the smallest possible value of the sum of squares of the residuals. To avoid the personal bias in fitting a curve to observed data, the method of least squares is used.

14.3.3 Scatter Diagram. The *scatter diagram* is a set of points in a rectangular co-ordinate system (with x measured horizontally and y measured vertically), where each point represents an observed pair of values. To aid in determining an equation connecting the two variables, a first step is the collection of the data showing the paired values of the variables under consideration.

Let us suppose that n pairs of observations (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) are made on two variables. The next step in the investigation is to plot the data on a graph to get a scatter diagram. The choice of the regression curve to fit may be influenced by theory, by experience or simply by looking at the scatter diagram. For example, the experiment may have been designed to verify a particular relationship between the variables. Alternately, the function form may be selected after inspecting the scatter diagram, as it would be useless to try to fit a straight line to some data if the relationship was clearly curvilinear. In practice the experimenter may choose the one which gives the best fit.

It is often possible to see, by looking at the scatter diagram that a smooth curve can be fitted to the data. In particular, if a straight line can be fitted to the data, then we say that a linear relationship exists between two variables, otherwise the relationship is curvilinear. A visual examination of a scatter diagram gives some useful indications of the nature and strength of the relationship between two variables and aids in choosing the appropriate type of model for estimation.

For example, if the points on the scatter diagram tend to run from the lower left side to the upper right side (that is, if the Y variable tends to increase as x increases), there is said to be a *direct* relationship between the two variables. On the other hand if the points on the scatter diagram tend to run from the upper left side to the lower right side (that is, if the variable Y tends to decrease as x increases), there is said to be *inverse* relationship between the two variables. The scatter diagram gives an indication whether a straight line appears to be an adequate

description of the average relationship between two variables. If a straight line is used to describe the average relationship between two variables, a linear relationship is said to exist. If the points on the scatter diagram appear to lie along a curve, a curvilinear relationship is said to be present.

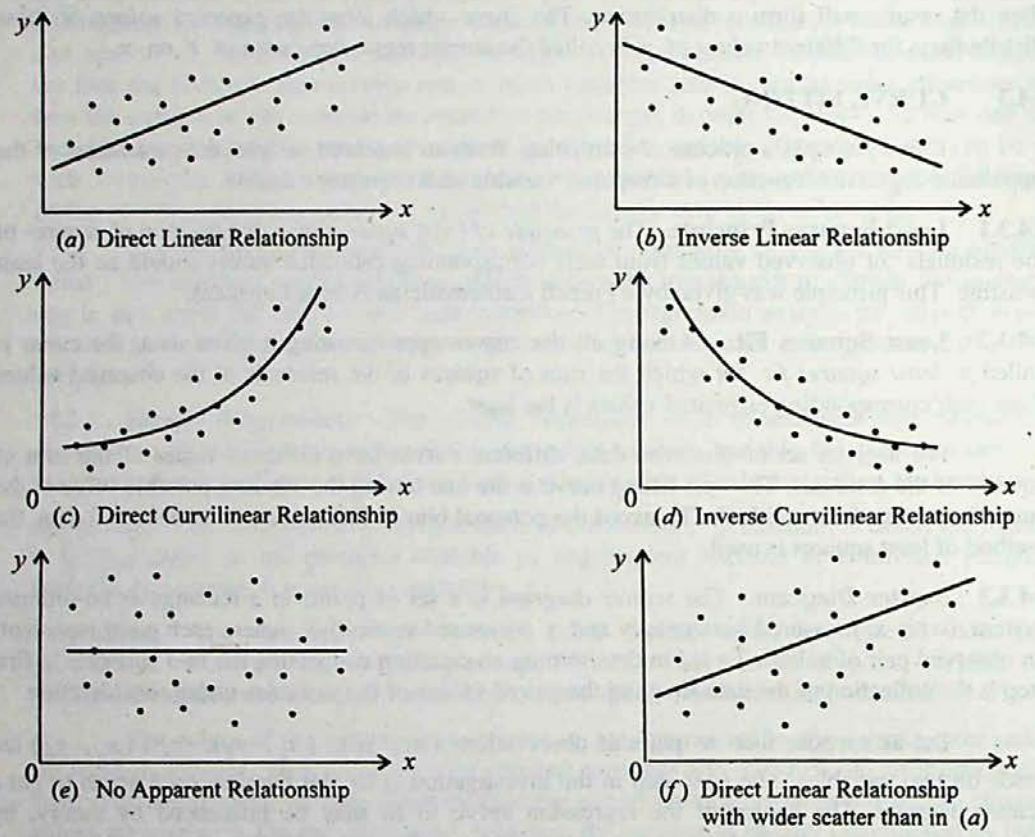


Fig. 14.1 Types of relationships found in scatter diagrams

Parts (*a*), (*b*), (*c*) and (*d*) of Fig 14.1 show direct linear, inverse linear, direct curvilinear and inverse curvilinear relationships. The points tend to follow a straight line with positive slope in (*a*), a straight line with negative slope in (*b*), a curve with positive slope in (*c*), and a curve with negative slope in (*d*). Of course the relationships are not always so obvious.

In (*e*) the points appear to follow a horizontal line. This type of scatter diagram depicts "no correlation" or no evident relationship between x and Y variables because the horizontal line implies no change, on the average, as x increases. In (*f*) the points follow a straight line with positive slope as in (*a*) but there is a much wider scatter of points around the line than in (*a*).

Note that a scatter diagram is primarily used to determine the appropriateness of a particular type of equation for describing the data. The approximate "goodness of fit" of the equation is also apparent from a scatter diagram, for example the fit in (*a*) is quite good as compared to the fit in (*f*). However, "goodness of fit" can and should be defined and determined precisely.

14.4 SIMPLE LINEAR REGRESSION

If the simple regression describes the dependence of the expected value of the dependent random variable Y as a linear function of the independent non-random variable x, then the regression is called *simple linear regression.* It is given by

$$\mu_{Y|x} = \alpha + \beta x$$

which implies that $\alpha = \mu_{Y|x}$ when $x = 0$. Thus α is the intercept of the line along *y-axis.* The β indicates the change in the mean of the probability distribution of Y per unit increase in x.

14.4.1 Simple Linear Regression Coefficient. The *simple linear regression coefficient* is the relative change in the expected value of the dependent random variable with respect to a unit increase in the independent non-random variable. It is denoted by β. The slope of the line β remains constant at each value of x.

It is measured by $\tan\theta$ where θ is the angle made by the line with the positive side of the *x-axis.* The slope of the line depends upon the value of the β. If the value of β is positive, the line will slope upward like the solid line in the Fig. 14.2. On the other hand, if the value of β is negative, the line will slope downward like the broken line in the Fig. 14.2.

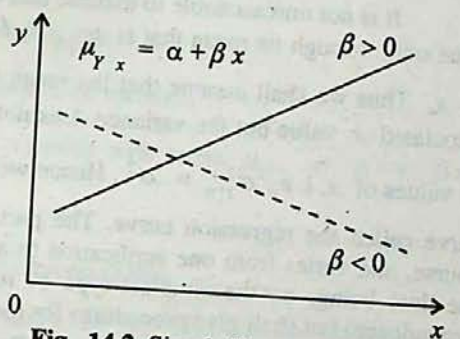


Fig. 14.2 Simple linear regression

14.5 THE SIMPLE LINEAR REGRESSION MODEL

We used the scatter diagram to illustrate the problem of selecting the regression line that provides the "best" estimate of the relationship between the independent and dependent variables in a regression problem. The next step is to specify the mathematical formulation of the linear regression model to provide a basis for statistical analysis.

14.5.1 An Example of Regression Model. In this section we will give an example to illustrate how the regression model between two variables can be used to simulate real world problems.

Suppose that we are to investigate the relationship between the consumption expenditure and the disposable income of households in a certain city for some given period of time. We know that as one's income increases, there is a tendency to spend more. What kind of relation is there between income and expenditure? Is it proportional, or is there any other form of a relationship, how close this relationship between income and expenditure is? Certainly there is no functional relationship between disposable income and consumption expenditure. Now let Y denote the consumption expenditure and x denote the disposable income.

Let us suppose that we have already divided our households into various groups on the basis of income levels. We do not expect that all the households within the group which have some given (fixed, predetermined) income x will display an identical expenditure. Some will spend more than the others, some will spend less, but we do expect a clustering of the expenditure figures around a central value with some variance. For each possible value of x chosen non-randomly there are several values of Y that could occur. Thus Y becomes a random

variable that possesses a distribution or population of associated y values for any given value of x. This distribution of associated y values, for any given x, is described either by a probability density function $f_Y(y|x)$ or by a probability mass function $p_Y(y|x)$ if the population has a discrete set of possible values. This distribution represents the relative likelihood of different values of Y occurring.

The mean of each probability distribution of Y values varies in some constant and systematic manner with the independent variable x. The mean of any distribution of Y for given x will be denoted by $\mu_{Y|x} = E(Y|x)$ and the variance of this distribution by $\sigma^2_{Y|x} = Var(Y|x)$. These are unknown parameters. They are constant for any fixed value of x but may vary between the distribution of Y_i for different x_i. The mean of Y for all values will be dented by μ_Y and the variance by σ_Y^2 or σ^2.

It is not unreasonable to assume that the random variable Y depends on the associated x value only through its mean that is $\mu_{Y|x} = f(x)$, but any higher moments of Y do not depend on x. Thus we shall assume that the mean value of the random variable Y depends upon the associated x value but the variance does not. We shall further assume that $\sigma^2_{Y|x}$ is constant for all values of x, *i. e.*, $\sigma^2_{Y|x} = \sigma^2$. Hence we assume that all the means $\mu_{Y|x}$ lie on a continuous curve called the regression curve. The particular form of the regression curve is arbitrary, of course, and varies from one application to another. We shall only concentrate our attention, for the time being, on the simplest type of regression curve, namely, the straight line (a linear dependence) but shall give procedures for more general models.

14.5.2 Mathematical Formulation of Regression Model. Our observed paired values of x and Y are only sample values from a large population. However, for a moment we are concerned with constructing a model for the population of all possible paired values. If, for example, a linear relationship is considered to be appropriate, that is, the average relationship between the dependent random variable Y and the independently varying non-random variable x is assumed to be linear. Since we are interested in the conditional expectation $\mu_{Y|x} = E(Y|x)$. By assuming that Y and x are linearly related, we are saying that all possible conditional means $\mu_{Y|x}$ which might be calculated one for each possible value of x must lie on a single straight line. This line is called the population regression line. To specify this line we need to know its slope and intercept. Let α be the y-intercept and β be the slope of the line. The population regression line is written as follows:

$$\mu_{Y|x} = \alpha + \beta x$$

This line is unknown. When some exact value of x is specified from its domain, it is customary to denote this value as x_i. Associated with this value x_i of the independent non-random variable x, there exists a random variable Y_i with a distribution or population with mean $\mu_{Y|x_i}$ and variance $\sigma^2_{Y|x_i}$. Assume that

$$\mu_{Y|x_i} = \alpha + \beta x_i, \qquad \sigma^2_{Y|x_i} = E(Y_i - \mu_{Y|x_i})^2 = \sigma^2$$

Now we define a deviation of the random variable Y_i from its unknown mean $\mu_{Y|x_i}$ and call this deviation as population regression error. For this reason, this difference is usually called the random "error" and denoted by ε_i. Therefore, we can define the random variable ε_i as

$$\varepsilon_i = Y_i - \mu_{Y|x_i} = Y_i - (\alpha + \beta x_i) = Y_i - \alpha - \beta x_i$$

There are three generally recognized sources of errors in such regression problems: (1) specification (or equation) error, arising from the omission of one or more relevant independent variables; (2) sampling error, arising from random variation of observations around their expected value and (3) measurement error, arising from the lack of precision in measuring variables. These errors are assumed to have zero mean and the constant variance identical to the variance of Y for a given value of x. We can now define what is called the population regression model as

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where, x_i = a predetermined value of a non-random variable.

Y_i = associated with x_i a random variable with mean $\mu_{Y|x_i} = \alpha + \beta x_i$ and variance $\sigma^2_{Y|x_i} = \sigma^2$.

α = the population y-intercept of the regression line.

β = the population slope of the regression line also known as the population regression coefficient.

ε_i = the deviation $(Y_i - \mu_{Y|x_i})$ in the population.

This model is said to be simple, linear in parameters and linear in independent variable. It is *simple*, in that there is only one independent variable, *linear* in the parameters because no parameter appears as an exponent or is multiplied or divided by another parameter, and *linear* in the independent variable, because the variable appears only in the first power.

14.5.3 The Sample Simple Linear Regression Model. In our population regression model α, β, $\mu_{Y|x}$ and σ^2 are unknown parameters we wish to estimate these parameters statistically on the basis of our sample observations on x and Y, and we may wish to test hypothesis and construct confidence intervals about these parameters. In this regard sampling is accomplished as follows:

(*i*) A set of n values of x in its domain is observed and denoted by $x_1, x_2, \cdots, x_n$. The x's are not random variables, but they may be selected either by some random procedure or by purposeful selection.

(*ii*) Each x_i determines a distribution or population whose mean is $\alpha + \beta x_i$ and whose variance is σ^2. From this distribution a value (a sample of size one) is selected at random and denoted by Y_i.

Thus we have a set of n pairs of observations denoted by $(Y_1, x_1), (Y_2, x_2), \cdots, (Y_n, x_n)$ which we have written to stress the fact that each sample of Y that we take has an associated x value. The values of x may or may not be all distinct, but as we shall see, we must have at least two different values of x represented if we are to estimate both α and β. We can write the n actually observed sample pairs as $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ without incorporating any further assumptions into our model, we can obtain estimates of α, β and $\mu_{Y|x}$. It is customary to let a be the best estimate of α, b be the best estimate of β and to let $\hat{y}$ be the resulting estimate of $\mu_{Y|x}$ and the line resulting a, b and $\hat{y}$ is called the best fitted regression line. This line has the same form as the population line. Thus the sample simple linear regression is

$$\hat{y} = a + bx$$

where $\hat{y}$ = the ordinate of the estimated line for any given value of x which is the best point estimate of $\mu_{Y|x}$

a = the y-intercept of the estimated line which is the best point estimate of α

b = the slope of the estimated line which is the best point estimate of β

Thus, if x_i is a specific value of x, then

$$\hat{y}_i = a + bx_i$$

is the equation for finding $\hat{y}_i$, which is the best estimate of $\mu_{Y|x_i}$ for this value x_i.

We can specify a sample regression model just as we did in the population regression model. Again we need to define an error term, which in this case is the deviation of actual value y_i from predicted value $\hat{y}_i$. This error term is denoted by e_i, which means that the sample regression error e_i is an estimate of the population error ε_i. The errors e_i ($i = 1, 2, \cdots, n$) are often called *residuals* or *deviations* or *prediction errors*.

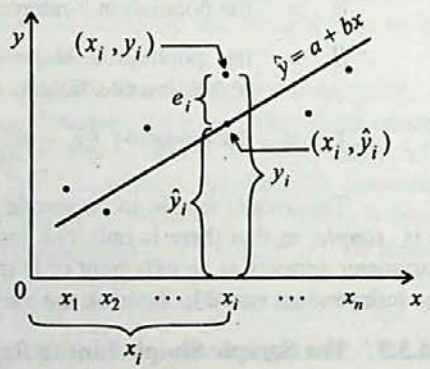


Fig. 14.3 Residual of y_i

Residual: $e_i = y_i - \hat{y}_i = y_i - (a + bx_i) = y_i - a - bx_i$

These residuals from the estimated line will be positive or negative as the actual value lies above or below the line. Thus the sample regression model is

$$y_i = a + bx_i + e_i$$

14.5.4 Covariance of Two Variables. If $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ are n pairs of observations on two variables X and Y, then the *covariance*, denoted by s_{xy}, is defined as

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n} \qquad i = 1, 2, \cdots, n$$

It is a measure of the *linear mutual variability* of the two variables. Its sign reflects the direction of the mutual variability: if the variables tend to move in the same direction, the covariance is positive; if the variables tend to move in opposite directions, the covariance is negative. It can be easily expressed as

$$s_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} \qquad i = 1, 2, \cdots, n$$

If $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$ are two series of n observations each, and if $z_i = x_i \pm y_i$, then

(*i*) $s_z^2 = s_x^2 + s_y^2 \pm 2 s_{xy}$

(*ii*) $s_z^2 = s_x^2 + s_y^2$ when X, Y are independent variables

14.5.5 Least Squares Point Estimation of α, β and $\mu_{Y|x_i}$ (Fitting of Straight Line). We now face the problem of estimating the linear regression between a dependent random variable Y and an independently varying non-random variable x given a sample of y values with their associated values of x. A general method of estimating the parameters of a regression line is the method of least squares which is explained in the following theorem.

***Theorem* 14.1** *Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be n observed values of a random variable Y, with their associated x values, where the regression line is $\mu_{Y|x} = \alpha + \beta x$.*

(*i*) *The least squares line is given by $\hat{y} = a + b\,x$, where*

$$b = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n \bar{x}\,\bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$a = \frac{\sum y_i - b \sum x_i}{n} = \bar{y} - b\bar{x}$$

(*ii*) *The least squares line always passes through the point of means $(\bar{x}, \bar{y})$.*

(*iii*) *The least squares estimate of $\mu_{Y|x_i}$ is*

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x})$$

***Example* 14.1** *The following sample of 8 grade point averages and marks in matriculation was observed for students from a college.*

Score	480	490	510	510	530	550	610	640
GPA	2.7	2.9	3.3	2.9	3.1	3.0	3.2	3.7

Find the least squares line. Estimate the mean GPA *of students scoring* 600 *marks.*

Solution. The estimated regression line is $\hat{y} = a + bx$

x_i	y_i	x_i^2	$x_i\, y_i$
480	2.7	230400	1296
490	2.9	240100	1421
510	3.3	260100	1683
510	2.9	260100	1479
530	3.1	280900	1643
550	3.0	302500	1650
610	3.2	372100	1952
640	3.7	409600	2368
4320	24.8	2355800	13492

The least squares estimates a and b are

$$b = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{8(13492) - (4320)(24.8)}{8(2355800) - (4320)^2} = 0.00435$$

$$a = \frac{\Sigma y_i - b\Sigma x_i}{n} = \frac{24.8 - 0.00435(4320)}{8} = 0.751$$

The best fitted line is $\hat{y} = 0.751 + 0.00435\, x$

For $x = 600$, we have $\hat{y} = 0.751 + 0.00435\,(600) = 3.361$

14.5.6 Properties of the Least Squares Line. The line fitted by the method of least squares has a number of properties worth noting.

(1) The sum of the residuals is zero, that is

$$\Sigma e_i = 0$$

However, rounding errors may, of course, be present in any particular case. Hence, in minimizing Σe_i^2 the least squares method automatically sets $\Sigma e_i = 0$.

(2) The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$

$$\Sigma y_i = \Sigma \hat{y}_i$$

Therefore, it follows that

(*i*) the mean of the fitted values $\hat{y}_i$ is the same as the mean of the observed values y_i

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{\Sigma \hat{y}_i}{n} = \bar{\hat{y}}_i$$

(*ii*) $\Sigma(\hat{y}_i - \bar{y})$ is also equal to zero.

(3) The sum of the squares of the residuals Σe_i^2 is minimum.

$$\Sigma e_i^2 = \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i\, y_i$$

(4) The regression line always passes through the point of means $(\bar{x}, \bar{y})$, the centre of gravity of the observed data. That is, whenever $x_i = \bar{x}$, we have $\hat{y}_i = \bar{y}$.

***Example* 14.2** *Given the following data*

x_i	0	1	2	3	4
y_i	1.0	1.8	3.3	4.5	6.3

(*a*) *Determine the least squares line taking x as independent variable.*

(*b*) *Find the estimated values for given values of x and show that*

(*i*) $\sum y_i = \sum \hat{y}_i$

(*ii*) $\sum e_i = 0$

(*c*) *Calculate the sum of the squares of the residuals.*

(*d*) *Verify that* $\sum e_i^2 = \sum y_i^2 - a\sum y_i - b\sum x_i\, y_i$.

Solution. (*a*) The estimated regression line is $\hat{y} = a + bx$

x_i	0	1	2	3	4	$\sum x_i = 10$
y_i	1.0	1.8	3.3	4.5	6.3	$\sum y_i = 16.9$
$x_i\, y_i$	0	1.8	6.6	13.5	25.2	$\sum x_i\, y_i = 47.1$
x_i^2	0	1	4	9	16	$\sum x_i^2 = 30$

The least squares estimates a and b are

$$b = \frac{n\sum x_i\, y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = \frac{5(47.1) - (10)(16.9)}{5(30) - (10)^2} = 1.33$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{16.9 - 1.33(10)}{5} = 0.72$$

The best fitted line is $\hat{y} = 0.72 + 1.33x$

(*b*) The estimated values $\hat{y}_i$ for the given values of x and the residuals $e_i = y_i - \hat{y}_i$ are obtained as shown in the following table.

x_i	y_i	$\hat{y}_i = 0.72 + 1.33x_i$	$e_i = y_i - \hat{y}_i$	e_i^2	y_i^2
0	1.0	0.72 + 1.33 (0) = 0.72	0.28	0.0784	1.00
1	1.8	0.72 + 1.33 (1) = 2.05	− 0.25	0.0625	3.24
2	3.3	0.72 + 1.33 (2) = 3.38	− 0.08	0.0064	10.89
3	4.5	0.72 + 1.33 (3) = 4.71	− 0.21	0.0441	20.25
4	6.3	0.72 + 1.33 (4) = 6.04	0.26	0.0676	39.69
Sum	16.9	16.90	0	0.2590	75.07

It is verified that

(*i*) $16.9 = \sum y_i = \sum \hat{y}_i = 16.9$

(*ii*) $\sum e_i = \sum (y_i - \hat{y}_i) = 0$

(*c*) The sum of the squares of the residuals is $\sum e_i^2 = 0.259$

(*d*) We are to verify that $\Sigma e_i^2 = \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i y_i$

$$0.259 = 75.07 - 0.72(16.9) - 1.33(47.1)$$

$$0.259 = 0.259$$

14.5.7 Coding and Scaling. In many cases the process of coding and scaling by a linear transformation can simplify the job of estimating the regression line or curve.

***Theorem* 14.2** *The sample linear regression coefficient b is independent of change of origin but it is not independent of change of scale.*

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be n observed values of a random variable Y, with their associated x values, then the sample regression coefficient is

$$b_{yx} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

Let $u_i = \dfrac{x_i - p}{h} \Rightarrow x_i = p + hu_i \Rightarrow \bar{x} = p + h\bar{u}$

and $v_i = \dfrac{y_i - q}{k} \Rightarrow y_i = q + kv_i \Rightarrow \bar{y} = q + k\bar{v}$

In this transformation we choose the constants p, q and h, k so that the transformed values u_i and v_i become as simple as possible. Then

$$b_{yx} = \frac{k}{h} b_{vu}$$

A Special Coding and Scaling. If the values $x_1, x_2, \cdots, x_n$ of the independent variable x are equally spaced at an interval h, then calculations involved in solving the normal equations can be made much simple by taking the origin at $\bar{x}$ and choosing a suitable unit of measurement. The choice of origin and unit is explained below in the two cases.

(*i*) If the sample size is odd, say $n = 2m - 1$, then we take the origin at the middle value x_m which is equal to $\bar{x}$, *i. e.*, $\bar{x} = x_m$. If h is the common interval, we take h as a new unit of measurement, then changing each x_i into u_i by a linear transformation $u_i = (x_i - \bar{x})/h$, the variable u takes the values $-(m-1), -(m-2), \cdots, -2, -1, 0, 1, 2, \cdots, (m-2), (m-1)$. Thus, we get

$$\Sigma u_i = 0 = \Sigma u_i^3 = \Sigma u_i^5 = \cdots$$

(*ii*) If the sample size is even, say $n = 2m$, then we take the origin at the average of the two middle values x_m and x_{m+1} which is equal to $\bar{x}$, *i.e.*, $\bar{x} = (x_m + x_{m+1})/2$. If h is the common interval, we take $h/2$ as a new unit of measurement, then changing each x_i into u_i by a linear transformation $u_i = (x_i - \bar{x})/(h/2)$, the variable u takes the values $-(2m-1), -(2m-3), \cdots, -3, -1, 1, 3, \cdots, (2m-3), (2m-1)$. Thus, we get

$$\Sigma u_i = 0 = \Sigma u_i^3 = \Sigma u_i^5 = \cdots$$

The values of the controlled variable are coded into integers symmetrically about zero. When the values of the coded variable u sum to zero, the least squares line of Y upon u becomes

$$\hat{y} = a + b\,u$$

where $a = \dfrac{\Sigma y_i}{n} = \bar{y}$ and $b = \dfrac{\Sigma u_i\, y_i}{\Sigma u_i^2}$

In the end, we must change the least squares line of Y on u into the least squares line of Y on x by transforming back the coded variable u into the original variable x. Sometimes, for the sake of further convenience each observed value y_i of the dependent variable can also be transformed into v_i by a linear transformation $v_i = (y_i - q)/k$ where q and k have arbitrary values.

***Example* 14.3** *The following table shows the tons of steel produced versus the number of workers in a small steel mill.*

Observation number i	*Number of workers* x_i	*Tons of steel produced* y_i
1	1	4
2	2	6
3	3	10
4	4	10
5	5	15
6	6	15
7	7	16
8	8	20

Estimate the line of regression using $u = \dfrac{x - 4.5}{1/2}$.

Solution. We have $\bar{x} = 4.5$ and $h = 1$. Let $u_i = \dfrac{x_i - \bar{x}}{h/2} = \dfrac{x_i - 4.5}{1/2}$

x_i	y_i	$u_i = \dfrac{x_i - 4.5}{1/2}$	$u_i\, y_i$	u_i^2
1	4	−7	− 28	49
2	6	−5	− 30	25
3	10	−3	− 30	9
4	10	−1	− 10	1
5	15	1	15	1
6	15	3	45	9
7	16	5	80	25
8	20	7	140	49
Sum	96	0	182	168

The estimated regression line of Y on u is $\hat{y} = a + b\,u$

The least squares estimates of a and b are

$$a = \frac{\Sigma y_i}{n} = \frac{96}{8} = 12$$

$$b = \frac{\Sigma u_i\, y_i}{\Sigma u_i^2} = \frac{182}{168} = 1.0833$$

The best fitted line of Y on u is $\hat{y} = 12 + 1.0833\,u$

Substituting $\dfrac{x - 4.5}{1/2}$ for u, we get the best fitted line of Y on x as

$$\hat{y} = 12 + 1.0833\left(\frac{x - 4.5}{0.5}\right) = 12 + \frac{1.0833}{0.5}(x - 4.5)$$

$$= 12 + 2.167\,(x - 4.5) = 12 + 2.167\,x - 9.75$$

$$= 2.25 + 2.167\,x$$

***Example* 14.4** *The following data show, in convenient units, the yield Y of a chemical reaction run at various different temperature x.*

$$n = 7, \quad \Sigma x_i = 980, \quad \Sigma y_i = 27.4, \quad \Sigma x_i\, y_i = 3958,$$

$$\Sigma x_i^2 = 140000, \quad \Sigma y_i^2 = 115.54$$

Assuming that a linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$ is appropriate estimate the regression line of yield on temperature. Find the residuals sum of squares.

Solution. The estimated regression line is $\hat{y} = a + b\,x$

The least squares estimates a and b are

$$b = \frac{n\,\Sigma x_i\, y_i - (\Sigma x_i)(\Sigma y_i)}{n\,\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{7(3958) - (980)(27.4)}{7(140000) - (980)^2} = 0.04357$$

$$a = \frac{\Sigma y_i - b\Sigma x_i}{n} = \frac{27.4 - 0.04357(980)}{7} = -2.1855$$

The best fitted line is $\hat{y} = -2.1855 + 0.04357\,x$

The sum of squares of the residuals is

$$\Sigma e_i^2 = \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i\, y_i$$

$$= 115.54 - (-2.1855)(27.4) - 0.04357\,(3958) = 2.97$$

14.5.8 Limitations of Linear Regression. There are a number of limitations and cautions that must be kept in mind when using linear regression. They are.

Firstly, the linear regression is applicable only to relationships that can be described by a straight line. Non-linear regression methods exist to deal with some non-linear relationships. If you are in doubt about whether the data are approximately linear, a scatter diagram will help you to decide.

Secondly, the procedure used to find the regression coefficients a and b will give us a linear equation which is the best fit (*i. e.*, has the lowest value of $\sum e_i^2$) for the data, even when a linear relationship is non-existent. Therefore, the test of significance must be made to determine whether the regression coefficient b is "real".

Thirdly, the regression equation predicts values of the dependent variable based on values of the independent variable. It is therefore an *asymmetrical* measure. The regression equation predicting Y based on x (called "regression Y on x") cannot be used to derive the equation that will predict x based on Y.

Finally, the regression equation holds only for the range of values actually observed. The regression equation will not necessarily hold beyond this range.

Exercise 14.1

1. (*a*) What is a scatter diagram? Describe its role in the theory of regression.

(*b*) Explain what is meant by

(*i*) regression, (*ii*) regressand, (*iii*) regressor

(*c*) Explain what is meant by

(*i*) simple linear regression, (*ii*) simple linear regression coefficient.

2. (*a*) The following measurements of the specific heat of a certain chemical were made in order to investigate the variation in specific heat with temperature.

Temperature ($^{\circ}C$)	x_i	0	10	20	30	40
Specific heat	y_i	0.51	0.55	0.57	0.59	0.63

Plot the points on a scatter diagram and verify that the relationship is approximately linear. Estimate the regression line of specific heat on temperature, and hence estimate the value of the specific heat when the temperature is $25\ ^{\circ}C$.
($\hat{y} = 0.514 + 0.0028\,x$; $\hat{y} = 0.584$)

(*b*) Determine the estimated regression equation $\hat{y} = a + b\,x$ in each of the following cases

(*i*) $n = 10,\ \sum x_i = 20,\ \sum y_i = 260,\ \sum x_i\,y_i = 3490,\ \sum x_i^2 = 3144$

(*ii*) $n = 100,\ \bar{x} = 125,\ \bar{y} = 80,\ \sum x_i\,y_i = 1007425,\ \sum x_i^2 = 1585000$

(*iii*) $\bar{x} = 52,\ \bar{y} = 237,\ \sum (x_i - \bar{x})^2 = 2800,\ \sum (x_i - \bar{x})(y_i - \bar{y}) = 9871$

(*iv*) $n = 8,\ \bar{x} = 7,\ \bar{y} = 5,\ \sum x_i\,y_i = 364,\ \sum (x_i - \bar{x})^2 = 132$

($\hat{y} = 24.0864 + 0.9568\,x$; $\hat{y} = 38.75 + 0.33\,x$; $\hat{y} = 53.70 + 3.525\,x$; $\hat{y} = 0.5459 + 0.6363\,x$)

3. (*a*) Estimate the regression line of Y on x for the following data.

x_i	25	30	35	40	45	50
y_i	78	70	65	58	48	42

Is it possible from the equation you have just found

(*i*) an estimate for the value of x when $y = 54$?

(*ii*) an estimate for the value of y, when $x = 37$? In each case, if the answer is "Yes", calculate the estimate. If the answer is "No", say why not.

{ $\hat{y} = 114.4 - 1.45x$; (*i*) No, x is controlled; (*ii*) 61 }

(*b*) From n pairs of values (x_i, y_i), $i = 1, 2, \cdots, n$ the following quantities are calculated

$$n = 20, \quad \Sigma x_i = 400, \quad \Sigma y_i = 220,$$
$$\Sigma x_i^2 = 8800, \quad \Sigma y_i^2 = 2620, \quad \Sigma x_i y_i = 4300$$

Find the linear regression equation of y on x and x on y. Which would be the more useful if.

(*i*) x is the age (in years) and y is the reaction time (in milliseconds) of 20 people;

(*ii*) x is the cost (in ,000 Rs.) and y the floor-space (in 100 ft ^{2}) of 20 buildings

{ $\hat{y} = 13.5 - 0.125x$; $\hat{x} = 25.5 - 0.5y$; (*i*) y on x; (*ii*) x on y }

4. (*a*) The following table shows the ages x and systolic blood pressures Y of 12 women.

Age (years) x_i	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure y_i	147	125	160	118	149	128	150	145	115	140	152	155

Assuming that a linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$ is appropriate, estimate the linear regression of blood pressure on age. Estimate the expected blood pressure of a woman whose age is 45 years. What is the change in blood pressure for a unit change in age?

($\hat{y} = 80.78 + 1.138x$; 132; 1.138)

(*b*) Suppose that four randomly chosen plots where treated with various levels of fertilizer, resulting in the following yields of corn.

Fertilizer (*kg*/Acre) x_i	100	200	400	500
Production (Bushels/ Acre) y_i	70	70	80	100

(*i*) Estimate the linear regression $\mu_{Y|x} = \alpha + \beta x$ of production Y on fertilizer x.

(*ii*) Estimate the yield when no fertilizer is applied.

(*iii*) Estimate the yield when the average amount of fertilizer is applied.

(*iv*) Estimate how much yield is increased for every kilogram of fertilizer applied.

{ (*i*) $\hat{y} = 59 + 0.070x$; (*ii*) when $x = 0$, $\hat{y} = 59$; (*iii*) when $x = \bar{x} = 300$, $\hat{y} = 80$; (*iv*) 0.070 bushels per kg of fertilizer }

5. (*a*) Describe the properties of the least squares regression line.

(*b*) Determine the regression line and estimate the weight of a student whose height is 68 inches.

Height (inches) x_i	72	66	67	69	74	61	66	62	70	63
Weight (pounds) y_i	178	141	158	165	180	133	159	140	160	136

Find also the estimated values for given values of height. Show that the sum of the estimated values is equal to the sum of the observed values of weight. Find the deviations $e_i = y_i - \hat{y}_i$. Show that these deviations add to zero.

($\hat{y} = -94.4 + 3.72x$; 158.76)

6. (*a*) Four identical money boxes contain different numbers of a particular type of coin and no coin of other types. The information on the combined weights, is given below.

Number of coins in box x_i	10	20	30	40
Combined weight of coins and box y_i	312	509	682	865

Estimate the regression line of Y on x. Estimate from your regression line,

(*i*) the weight of an empty box,

(*ii*) the mean weight of a single coin. State the co-ordinates of one point through which the line of regression of Y upon x must pass.

{ $\hat{y} = 134 + 18.32x$; (*i*) 134, (*ii*) 18.32; (25, 592) }

(*b*) Fifteen boys took two examination papers in the same subject and their marks as percentages were as follows, where each boy's marks are in the same column.

Paper I x_i	65	73	42	52	84	60	70	79	60	83	57	77	54	66	89
Paper II y_i	78	88	60	73	92	77	84	89	70	89	73	88	70	85	89

Calculate the equation of the line of regression of Y on x. Two boys were each absent from one paper. One scored 63 on paper I, the other scored 81 on paper II. In which case can you use your regression line to estimate the mark that the boy should be allocated for the paper he did not take, and what is that mark ?

($\hat{y} = 35.53 + 0.665x$, 63 on I $\Rightarrow$ 78 on II)

7. (*a*) The following data shows the son's height and father's height.

Father's height (inches) x_i	59	61	63	65	67	69	71	73	75
Son's height (inches) y_i	64	66	67	67	68	69	70	72	72

Estimate the regression line $\mu_{Y|x} = \alpha + \beta x$ of son's height on father's height using $u_i = (x_i - 67)/2$ and $v_i = y_i - 68$. Predict the mean height of sons whose fathers are 70 inches in height.

($\hat{y} = 35.95 + 0.4833x$; 69.78)

(*b*) For 9 observations on supply X and price Y the following data was obtained

$\Sigma(x_i - 90) = -25$, $\Sigma(x_i - 90)^2 = 301$, $\Sigma(y_i - 127) = 12$,

$\Sigma(y_i - 127)^2 = 1006$, $\Sigma(x_i - 90)(y_i - 127) = -469$

Obtain the estimated line of regression of X on Y and estimate the supply when the

price is Rs. 125.
($\hat{x}$ = 143.69 − 0.44 y; 88.69)

(c) Number of revolutions x (per minute) and power y (hp) of a diesel engine are

x_i	400	500	600	700	800
y_i	580	1030	1420	1880	2310

Determine the regression line of the y-values on the x – values of the sample using $x_i = 100\,u_i + 600$ and $y_i = 10\,v_i + 1400$ estimate y when $x = 750$.
($\hat{y}$ = − 1142 + 4.31 x; 2090.5)

8. (a) Fit a straight line taking x as independent variable

$3x_i + 2$	5	8	11	14	17	20	23	26
$3y_i - 2$	7	10	16	16	25	28	28	34

Also estimate y for $x = 5/3$.
($\hat{y}$ = 1.714 + 1.286 x; 3.86)

(b) Fit a least squares line to following data taking (*i*) Y as dependent variable (*ii*) X as dependent variable.

x_i	1	3	4	6	8	9	11	14
y_i	1	2	4	4	5	7	8	9

Show that the two least squares lines obtained intersect at the point $(\bar{x}, \bar{y})$. estimate the mean value of y when $x = 7$. Estimate the mean value of x when $y = 6$.
($\hat{y}$ = 0.5455 + 0.6364 x, $\hat{y}$ = 5 for x = 7;
$\hat{x}$ = −0.5 + 1.5 y, $\hat{x}$ = 8.5 for y = 6)

9. (a) A random sample of 5 pairs of observations (x_i, y_i) is given below

x_i	3	2	5	1	4
y_i	13	9	27	8	18

Determine the least squares linear regression $\hat{y} = a_{yx} + b_{yx}\,x$ and estimate y for $x = 6$. Also find the least squares linear regression $\hat{x} = a_{xy} + b_{xy}\,y$ and use this to find that value of y for which $\hat{x} = 6$. Account for the difference.
($\hat{y}$ = 0.9 + 4.7 x, $\hat{y}$ = 29.1 for x = 6; $\hat{x}$ = 0.09 + 0.194 y, y = 30.46 for $\hat{x}$ = 6 which is a useless estimate, because the regression analysis does not permit the inverse use of the least squares line.)

(b) Compute the regression coefficients in each of the following cases:

(*i*) $n = 10$, $\Sigma(x_i - \bar{x})^2 = 170$, $\Sigma(y_i - \bar{y})^2 = 140$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 92$

(*ii*) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 148$, $s_x = 7.933$, $s_y = 16.627$, $n = 15$

($b_{yx} = 0.54$, $b_{xy} = 0.66$; $b_{yx} = 0.16$, $b_{xy} = 0.04$)

◆

14.6 SIMPLE LINEAR CORRELATION

The *simple linear correlation* measures the strength or closeness of linear relationships between two variables. The purpose of simple linear correlation is to determine whether or not two variables are related, that is, whether one variable tends to increase (or decrease) as the other variable increases. The correlation analysis is performed keeping in view the following two aspects.

(*i*) It measures the closeness of the linear regression to the distribution of observations of a dependent variable with associated values of an independent variable.

(*ii*) It measures the degree (extent or strength) of covariability between two variables.

We have discussed this first aspect in the preceding chapter. We shall now discuss the second aspect of correlation. This approach to the problem of understanding the relationship between two variables is to leave the type or form of the relationship unspecified and concentrate on measuring the strength of the relationship itself.

14.6.1 Positive Correlation. The correlation is said to be *positive* (or *direct*) if the two random variables tend to move in the same direction, *i. e.*, increase (or decrease) simultaneously. That is, the correlation is positive if the least squares regression lines have positive slopes.

Perfect Positive Correlation. The correlation is said to be *perfect positive* if the relationship between the two random variables is perfectly linear with positive slope.

14.6.2 Negative Correlation. The correlation is said to be *negative* (or *inverse*) if the two random variables tend to move in opposite directions, *i. e.*, one random variable decreases as the other random variable increases. That is, the correlation is negative if the least squares regression lines have negative slopes.

Perfect Negative Correlation. The correlation is said to be *perfect negative* if the relationship between the two random variables is perfectly linear with negative slope.

14.6.3 No Correlation. If one least squares regression line is horizontal and the other least squares regression line is vertical then there is no correlation between the two random variables. That is, if X and Y are independent, then $Cov(X, Y) = 0$ which implies that $\rho = 0$ and we say that there is no correlation.

14.7 CORRELATION ANALYSIS

One of the most widely used statistical techniques applied by statistician is *correlation analysis.* In purely correlation problems both the variables X and Y are random and the relationship between them is considered simultaneously and symmetrically. Examples of correlation problems are: (*i*) heights and weights of persons, (*ii*) ages of husbands and ages of wives at the time of their marriages, (*iii*) I. Q. of brothers and I. Q. of sisters, (*iv*) marks of students in economics and in statistics, (*v*) income and I. Q. of persons, (*vi*) demand and supply of a commodity, (*vii*) daily wages and overtime wages, (*viii*) gold prices and silver prices, (*ix*) the height and the circumference of head of babies at the time of their birth, (*x*) the greatest and the smallest diameters of hen eggs, *etc.*

In correlation problems, we sample from a population, observing two measurements on each individual in the sample. For example, if a person is selected at random, and both his height and weight are left free to take any possible values. Thus we have a joint distribution of two random variables or we may say that we have bivariate distribution. The data are assumed to be obtained by taking a random sample of values of X and Y.

14.7.1 Sample Correlation Coefficient. If $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ is a random sample of n pairs of observations from a bivariate population, then the sample correlation coefficient, denoted by r or more appropriately r_{xy}, is defined as

$$r = \frac{s_{xy}}{s_x s_y}$$

It can be expressed as

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\{\Sigma(x_i - \bar{x})^2/n\}\{\Sigma(y_i - \bar{y})^2/n\}}}$$

$$= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)(\Sigma y_i^2 - n\bar{y}^2)}}$$

$$= \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{\{\Sigma x_i^2 - (\Sigma x_i)^2/n\}\{\Sigma y_i^2 - (\Sigma y_i)^2/n\}}},$$

$$= \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\{n\Sigma x_i^2 - (\Sigma x_i)^2\}\{n\Sigma y_i^2 - (\Sigma y_i)^2\}}}$$

This r is the maximum likelihood estimate of ρ. The process of subtracting $\bar{x}$ and $\bar{y}$ indicates that the origin has been shifted to $(\bar{x}, \bar{y})$.

14.7.2 Properties of Sample Correlation Coefficient r. The sample correlation coefficient r has the following properties.

(1) r is symmetrical with respect to the variables X and Y, that is

$$r_{xy} = r_{yx}$$

(2) r is the covariance of values of the two variables X and Y measured in standard units, that is

$$r = Cov(z_x, z_y)$$

(3) ***Change of Origin and Scale.*** The value of r remains unchanged if constants are added to or subtracted from the values of the variables or if the values of the variables are multiplied or divided by constants having the same sign, but the value of r changes in sign only if the values of the variables are multiplied or divided by constants having opposite signs. That is, the magnitude of the sample correlation coefficient $|r|$ is independent of change of origin and scale.

(4) r always lies between -1 and $+1$, *i.e.*,

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

(5) $|r|$ is the geometric mean of the two regression coefficients b_{yx} and b_{xy}, that is

$$r = (+/-)\sqrt{b_{yx} \times b_{xy}}.$$

$$\text{Thus} \quad r = \begin{cases} +\sqrt{b_{yx} \times b_{xy}} & \text{if } b_{yx} \text{ and } b_{xy} \text{ are positive} \\ -\sqrt{b_{yx} \times b_{xy}} & \text{if } b_{yx} \text{ and } b_{xy} \text{ are negative} \end{cases}$$

(6) r is zero when one of the variables X or Y is constant.

***Theorem* 14.3** *The correlation coefficient is independent of the origin and the scale of measurement of the variables.*

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ *be a random sample of* n *pairs of observations from a bivariate population.*

If we let $u_i = (x_i - p)/h$ *and* $v_i = (y_i - q)/k$ *then* $r_{xy} = \dfrac{hk}{|h||k|} r_{uv}$

$$\text{That is,} \quad r_{xy} = \begin{cases} r_{uv} & \text{if } h \text{ and } k \text{ have same signs} \\ -r_{uv} & \text{if } h \text{ and } k \text{ have different signs} \end{cases}$$

Similarly, if we let $u_i = p + h x_i$ *and* $v_i = q + k y_i$, *then* $r_{uv} = \dfrac{hk}{|h||k|} r_{xy}$

$$\text{That is,} \quad r_{uv} = \begin{cases} r_{xy} & \text{if } h \text{ and } k \text{ have same signs} \\ -r_{xy} & \text{if } h \text{ and } k \text{ have different signs} \end{cases}$$

***Example* 14.5** *The following are the measurements of height and weight of* 8 *men.*

Height (inches) x_i	78	89	97	69	59	79	68	61
Weight (pound) y_i	125	137	156	112	107	136	123	104

(*i*) *Calculate the correlation coefficient between the height and weight of eight men by using the deviations from their means.*

(*ii*) *Again compute the correlation coefficient by taking the deviations of variable* X *from* 70 *and of variable* Y *from* 120.

(*iii*) *Do the results in* (*i*) *and* (*ii*) *agree?*

Solution. (*i*) The coefficient of correlation between X and Y is

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
78	125	3	0	9	0	0
89	137	14	12	196	144	168
97	156	22	31	484	961	682
69	112	−6	−13	36	169	78
59	107	−16	−18	256	324	288
79	136	4	11	16	121	44
68	123	−7	−2	49	4	14
61	104	−14	−21	196	441	294
600	1000	0	0	1242	2164	1568

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{600}{8} = 75$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1000}{8} = 125$$

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}} = \frac{1568}{\sqrt{(1242)(2164)}} = 0.956$$

(*ii*) Let the assumed mean for x be $p = 70$, and $u_i = x_i - p = x_i - 70$. Let the assumed mean for y be $q = 120$, and $v_i = y_i - q = y_i - 120$.

x_i	y_i	$u_i = x_i - 70$	$v_i = y_i - 120$	u_i^2	v_i^2	$u_i v_i$
78	125	8	5	64	25	40
89	137	19	17	361	289	323
97	156	27	36	729	1296	972
69	112	−1	−8	1	64	8
59	107	−11	−13	121	169	143
79	136	9	16	81	256	144
68	123	−2	3	4	9	−6
61	104	−9	−16	81	256	144
		40	40	1442	2364	1768

The coefficient of correlation between U and V is

$$r_{uv} = \frac{n \Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{\{n \Sigma u_i^2 - (\Sigma u_i)^2\}\{n \Sigma v_i^2 - (\Sigma v_i)^2\}}}$$

$$= \frac{8(1768) - (40)(40)}{\sqrt{\{8(1442) - (40)^2\}\{8(2364) - (40)^2\}}} = 0.956$$

(*iii*) The results in (*i*) and (*ii*) are same, since $0.956 = r_{xy} = r_{uv} = 0.956$.

***Example* 14.6** The following data were obtained for a sample of 10 persons from a height and weight distribution.

$$\Sigma x_i = 700, \quad \Sigma y_i = 1550, \quad \Sigma x_i^2 = 49120, \quad \Sigma y_i^2 = 240550, \quad \Sigma x_i y_i = 108650$$

Compute the coefficient of correlation.

Solution. The coefficient of correlation between X and Y is

$$r = \frac{n \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\{n \Sigma x_i^2 - (\Sigma x_i)^2\}\{n \Sigma y_i^2 - (\Sigma y_i)^2\}}}$$

$$= \frac{10(108650) - (700)(1550)}{\sqrt{\{10(49120) - (700)^2\}\{10(240550) - (1550)^2\}}} = 0.79$$

14.7.3 Goodness of Fit of a Linear Regression Equation. One approach to correlation analysis emphasized the covariability of the two random variables. The other approach to correlation analysis is related to regression analysis and provides a measure of the strength of closeness of the linear relationship between two variables; thus correlation coefficient is a measure of the goodness of fit of the linear regression equation. Consider a sample from a bivariate distribution of X and Y. There are two regression functions, each obtained by considering that variable as dependent whose mean value is to be estimated and treating the other variable as independent. The two linear regression functions of Y on X and of X on Y are

$$\mu_{Y|x} = \alpha_{YX} + \beta_{YX} X, \qquad \mu_{X|y} = \alpha_{XY} + \beta_{XY} Y$$

where β_{YX} and β_{XY} are the population regression coefficients of Y on X and of X on Y respectively. Their corresponding least squares sample regression lines are

$$\hat{y} = a_{yx} + b_{yx}x \qquad \hat{x} = a_{xy} + b_{xy}y$$

where b_{yx} and b_{xy} are the sample regression coefficients of Y on X and of X on Y respectively. The least squares estimates are

$$b_{yx} = \frac{s_{xy}}{s_x^2}, \qquad a_{yx} = \bar{y} - b_{yx}\bar{x}$$

$$b_{xy} = \frac{s_{xy}}{s_y^2}, \qquad a_{xy} = \bar{x} - b_{xy}\bar{y}$$

The regression equation of Y on X becomes

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

and the regression equation of X on Y becomes

$$\hat{x} = \bar{x} + b_{xy}(y - \bar{y})$$

The regression coefficients are related to the correlation coefficient as

$$b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{r\,s_y}{s_x}, \qquad b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{r\,s_x}{s_y}$$

Thus the regression equation of Y on X becomes

$$\hat{y} = \bar{y} + \frac{r\,s_y}{s_x}(x - \bar{x})$$

and the regression equation of X on Y becomes

$$\hat{x} = \bar{x} + \frac{r\,s_x}{s_y}(y - \bar{y})$$

Since s_x and s_y are positive, the sign of s_{xy}, b_{yx}, b_{xy} and r_{xy} will always be same. Note also that if any one of these four quantities is zero then all others must equal to zero. A positive sign of r_{xy} indicates that X and Y are directly related. A direct relationship between X and Y is associated with an upward sloping regression line; that is as one variable increases other variable also increases. A negative sign of r_{xy} indicates that X and Y are inversely related. An inverse relationship between X and Y is associated with a downward sloping regression line, that is as one variable increases, the other variable decreases.

***Theorem* 14.4** *In the correlation analysis the two regression lines intersect at the* point $(\bar{x}, \bar{y})$.

***Theorem* 14.5** *The correlation coefficient r is the slope of the regression lines for standard scores.*

***Theorem* 14.6** *The graphs of the regression lines of Y on X and X on Y are identical if all the points of the given sample lie on a straight line.*

***Example* 14.7** *The following data were obtained for a sample of* 10 *men from a height and weight distribution.*

$$\bar{x} = 70, \qquad \bar{y} = 155, \qquad \sum (x_i - \bar{x})^2 = 120,$$

$$\sum y_i^2 = 240550, \qquad \sum (x_i - \bar{x})(y_i - \bar{y}) = 150$$

Calculate covariance, correlation coefficient, the two regression lines.

Solution. The variance of X, variance of Y, covariance and correlation coefficient are

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{120}{10} = 12$$

$$s_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2 = \frac{240550}{10} - (155)^2 = 30$$

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{150}{10} = 15$$

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{15}{\sqrt{(12)(30)}} = 0.79$$

The estimated regression line of Y on X is $\hat{y} = a_{yx} + b_{yx} x$

The least squares estimates of a_{yx} and b_{yx} are

$$b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{15}{12} = 1.25$$

$$a_{yx} = \bar{y} - b_{yx} \bar{x} = 155 - 1.25(70) = 67.5$$

The best fitted line of Y on X is $\hat{y} = 67.5 + 1.25 x$

The estimated regression line of X on Y is $\hat{x} = a_{xy} + b_{xy} y$

The least squares estimates of a_{xy} and b_{xy} are

$$b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{15}{30} = 0.50$$

$$a_{xy} = \bar{x} - b_{xy} \bar{y} = 70 - (0.50)(155) = -7.5$$

The best fitted line of X on Y is $\hat{x} = -7.5 + 0.5 y$

***Example* 14.8** *Find the coefficient of correlation if the two regression coefficients have the following values*

(*i*) 0.45 and 0.8, (*ii*) -0.1 and -0.4.

Solution.

(*i*) $r = (+/-) \sqrt{b_{yx} \times b_{xy}} = + \sqrt{(0.45)(0.8)} = 0.6$

(*ii*) $r = (+/-) \sqrt{b_{yx} \times b_{xy}} = - \sqrt{(-0.1)(-0.4)} = -0.2$

***Example* 14.9** *The coefficient of correlation, for a sample of* 20 *pairs of observations is* 0.6. *If* $\bar{x} = 12$, $\bar{y} = 20$, $s_x = 1.5$ and $s_y = 2$, *estimate the lines of the regression. Estimate the mean of* Y *for* $x = 10$. *Estimate the mean of* X *for* $y = 22$.

Solution. The estimated regression line of Y on X is $\hat{y} = a_{yx} + b_{yx}\, x$

The least squares estimates of a_{yx} and b_{yx} are

$$b_{yx} = \frac{r\, s_y}{s_x} = \frac{0.6(2)}{1.5} = 0.8$$

$$a_{yx} = \bar{y} - b_{yx}\, \bar{x} = 20 - 0.8\,(12) = 10.4$$

The best fitted line of Y on X is

$$\hat{y} = 10.4 + 0.8\, x$$

For $x = 10$, we have

$$\hat{y} = 10.4 + 0.8\,(10) = 18.4$$

The estimated regression line of X on Y is $\hat{x} = a_{xy} + b_{xy}\, y$

The least squares estimates of a_{xy} and b_{xy} are

$$b_{xy} = \frac{r\, s_x}{s_y} = \frac{0.6(1.5)}{2} = 0.45$$

$$a_{xy} = \bar{x} - b_{xy}\, \bar{y} = 12 - 0.45\,(20) = 3$$

The best fitted line of X on Y is

$$\hat{x} = 3 + 0.45\, y$$

For $y = 22$, we have

$$\hat{x} = 3 + 0.45\,(22) = 12.9$$

***Example* 14.10** *The following results are given from paired data of two variables* X *and* Y.

Estimate of variance of $X = 9$

Estimated regression line of X *on* Y: $40\,\hat{x} - 18\, y = 214$

Estimated regression line of Y *on* X: $8\, x - 10\,\hat{y} = -66$

Find (i) *The coefficient of correlation between* X *and* Y, (ii) *Standard deviation of* Y, (iii) *Mean values of* X *and* Y.

Solution. (i) The estimated regression line of X on Y is

$$40\hat{x} - 18\, y = 214 \quad \Rightarrow \quad \hat{x} = \frac{214}{40} + \frac{18}{40} y \quad \Rightarrow \quad b_{xy} = \frac{18}{40} = 0.45$$

The estimated regression line of Y on X is

$$8x - 10\hat{y} = -66 \quad \Rightarrow \quad \hat{y} = \frac{66}{10} + \frac{8}{10}x \quad \Rightarrow \quad b_{yx} = \frac{8}{10} = 0.8$$

The estimate of the correlation coefficient between X on Y is

$$r = (+/-)\sqrt{b_{yx} \times b_{xy}} = +\sqrt{(0.8)(0.45)} = 0.6$$

(ii) $s_x^2 = 9 \quad \Rightarrow \quad s_x = +\sqrt{9} = 3$

$$b_{yx} = \frac{r_{xy}\, s_y}{s_x}$$

$$0.8 = \frac{0.6\, s_y}{3} \quad \Rightarrow \quad 0.6\, s_y = 0.8(3) \quad \Rightarrow \quad s_y = 4$$

(iii) Since both the estimated regression lines pass through the point $(\bar{x}, \bar{y})$. Thus

$$40\bar{x} - 18\bar{y} = 214 \ldots\ldots (i)$$

$$8\bar{x} - 10\bar{y} = -66 \ldots\ldots (ii)$$

Multiplying *(ii)* by 5 and subtracting it from *(i)*

$$\begin{array}{rrr} 40\bar{x} - 18\bar{y} = & 214 \\ 40\bar{x} - 50\bar{y} = & -330 \\ - \quad + & + \\ \hline 32\bar{y} = & 544 \end{array} \quad \Rightarrow \quad \bar{y} = 17$$

Putting this value of $\bar{y}$ in *(ii)*, we have

$$8\bar{x} - 10(17) = -66 \quad \Rightarrow \quad \bar{x} = 13$$

14.7.4 Correlation and Causation. It is necessary to consider the sampling distribution of the sample statistic R to decide whether or not we should accept the hypothesis that the variables in the population are related. But aside from this technical aspect of a relation between two variables, it is necessary for a statistician to consider whether or not correlation indicates a cause and effect relationship. It is possible to correlate the temperature of Lahore city with the birth rate and it is possible that a high positive correlation may be found showing that when the temperature is high, the birth rate is high, and when the temperature is low the birth rate is low.

There is no meaning to such a correlation. There is no causal relationship between the two phenomena. This example illustrate that you can correlate anything, and there are chances you may obtain a high correlation which may have no significant meaning at all. A high correlation simply tells us that the data we have collected is *consistent* with the hypothesis we set up. That is, it supports our hypothesis. We may say the following situations that brought about a high correlation.

(*i*) X is the cause of Y.

(*ii*) Y is the cause of X.

(*iii*) There is a third factor Z that affects X and Y such that they show a close relation.

(*iv*) The correlation between X and Y may be due to chance.

Only by more thorough investigation we can come to some conclusion as to whether or not X is the cause of Y.

Exercise 14.2

1. **(a)** Differentiate between regression and correlation problems, giving examples.

(b) Define the terms correlation and product moment co-efficient of correlation.

(c) For a set of 50 pairs of observations on variables X and Y, we have $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 450$. Find the covariance.
$(s_{xy} = 9)$

2. **(a)** The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}}$$

The following table gives the ages of husbands and the ages of wives at the time of their marriage.

Couple	i	1	2	3	4	5	6	7	8	9	10
Husband's age	x_i	25	29	30	30	31	32	33	35	37	38
Wife's age	y_i	20	22	24	29	23	31	29	31	30	31

Calculate the coefficient of correlation by using the above formula.
$(r = 0.82)$

(b) The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{\{\Sigma x_i^2 - n\bar{x}^2\}\{\Sigma y_i^2 - n\bar{y}^2\}}}$$

The following table gives the demand and supply of a commodity.

Supply	x_i	400	200	700	100	500	300	600
Demand	y_i	50	60	20	70	40	30	10

Calculate the coefficient of correlation by using the above formula.
$(r = -0.857)$

(c) The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{\{\Sigma x_i^2 - (\Sigma x_i)^2/n\}\{\Sigma y_i^2 - (\Sigma y_i)^2/n\}}}$$

The following table gives the traffic density and accident rate.

Traffic density	x_i	30	35	40	45	50	60	70	80	90
Accident rate	y_i	2	4	5	5	8	15	24	30	32

Calculate the coefficient of correlation by using the above formula.
($r = 0.983$)

(d) The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\{n\sum x_i^2 - (\sum x_i)^2\}\{n\sum y_i^2 - (\sum y_i)^2\}}}$$

The following table gives the number of persons employed and cloth manufactured in a textile mill..

Persons employed	x_i	137	209	113	189	176	200	219
Cloth manufactured	y_i	23	47	22	40	39	51	49

Calculate the coefficient of correlation by using the above formula.
($r = 0.963$)

3. (a) For a set of 22 pairs of observations, we have

$$\sum x_i = 983, \quad \sum y_i = 409, \quad \sum x_i^2 = 61339, \quad \sum y_i^2 = 8475, \quad \sum x_i y_i = 15811$$

Find the product moment correlation coefficient for the data.
($r = -0.6325$)

(b) For a sample of 20 pairs of observations, we have

$$\bar{x} = 2, \quad \bar{y} = 8, \quad \sum x_i^2 = 180, \quad \sum y_i^2 = 3424, \quad \sum x_i y_i = 604$$

Calculate the coefficient of correlation.
($r = 0.6133$)

(c) For a set of 8 pairs of observations, we have

$$\sum x_i = 448, \quad \sum y_i = 472, \quad \sum y_i^2 = 29958, \quad \sum x_i y_i = 26762, \quad s_x = 16.6$$

Compute the product moment correlation coefficient.
($r = 0.15$)

4. (a) For a set of 50 pairs of observations, the standard deviations of x and y are 4.5 and 3.5 respectively. If the sum of products of deviations of x and y values from their respective means be 420, find the Karl Pearson's coefficient of correlation.
($r = 0.53$)

(b) For a given set of data, we have $s_x^2 = 9.102, \quad s_y^2 = 2.204, \quad s_{xy} = 1.694$

Find the product moment correlation coefficient for the data.
($r = 0.378$)

5. (a) For a given set of data, we have $r = 0.48, \quad s_{xy} = 36, \quad s_x^2 = 16$. Find s_y.
($s_y = 18.75$)

(b) For a given set of data, we have

$$r = 0.5, \quad \sum(x_i - \bar{x})(y_i - \bar{y}) = 120, \quad s_y = 8, \quad \sum(x_i - \bar{x})^2 = 90$$

Find the number of pairs of values.
($n = 10$)

6. (a) A computer while calculating the correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results.

$$\sum x_i = 125, \quad \sum x_i^2 = 650, \quad \sum y_i = 100, \quad \sum y_i^2 = 460, \quad \sum x_i y_i = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as:

x_i	6	8
y_i	14	6

while the correct values were:

x_i	8	6
y_i	12	8

Obtain the correct value of coefficient of correlation.
($r = 0.67$)

(*b*) The following data show the marks in economics and marks in statistics obtained by ten students.

Student	i	1	2	3	4	5	6	7	8	9	10
Economics	x_i	78	36	96	25	75	82	90	62	65	39
Statistics	y_i	84	51	91	60	68	62	86	58	53	47

(*i*) Compute the coefficient of correlation.

(*ii*) Again compute the coefficient of correlation by taking the deviations of variable X from 50 and of variable Y from 60.

(*iii*) Do the results in (*i*) and (*ii*) agree?

{ (*i*) 0.775, (*ii*) 0.775, (*iii*) Yes }

(*c*) Compute the correlation coefficient between the variables X and Y represented in the following table:

x_i	2	4	5	6	8	11
y_i	18	12	10	8	7	5

Multiply each x_i value by 2 and add 6. Multiply each y_i value by 3 and subtract 15. Find the correlation co-efficient between the two new sets of values, explaining why you do or do not obtain the same result as above.
(– 0.92)

7. (*a*) Interpret the meaning when

$r = -1$, $r = 0$, $r = 1$

(*b*) Sketch scatter diagrams which illustrate:

(*i*) positive linear correlation, (*ii*) perfect positive linear correlation,

(*iii*) negative linear correlation, (*iv*) perfect negative linear correlation,

(*v*) no correlation, between two variables X and Y.

8. (*a*) From the data given below, calculate the coefficient of correlation between the ages of husbands and ages of wives at the time of their marriage.

Couple	i	1	2	3	4	5	6	7	8	9	10
Husband's age	x_i	28	27	28	23	29	30	36	35	33	31
Wife's age	y_i	27	20	22	18	21	29	29	28	29	27

Find the regression coefficients. Verify that r is the geometric mean of the two regression coefficients
($r = 0.82$, $b_{yx} = 0.89$, $b_{xy} = 0.75$)

(*b*) The two regression coefficients have following values, find r.

(*i*) $b_{yx} = 0.86,\ b_{xy} = 0.95$

(*ii*) $b_{yx} = -0.52,\ b_{xy} = -1.02$

{ (*i*) $r = 0.90$, (*ii*) $r = -0.73$ }

(*c*) Find the two regression coefficients in each of the following cases.

(*i*) $\Sigma x_i = 17.6,\quad \Sigma y_i = 32.8,\quad \Sigma x_i y_i = 94.7,$
$\Sigma x_i^2 = 49.64,\quad \Sigma y_i^2 = 182,\quad n = 8$

(*ii*) $n = 10,\quad \Sigma(x_i - \bar{x})^2 = 170,\quad \Sigma(y_i - \bar{y})^2 = 140,$
$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 92$

(*iii*) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 148,\quad s_x = 7.933,\quad s_y = 16.627,\quad n = 15$

(*iv*) $n = 8,\quad \bar{x} = 7,\quad \bar{y} = 5,\quad \Sigma x_i y_i = 364,$
$\Sigma(x_i - \bar{x})^2 = 132,\quad \Sigma(y_i - \bar{y})^2 = 56.$

(*v*) $r = 0.97,\quad s_x = 17.08,\quad s_y = 14.34$

{ (*i*) $b_{yx} = 2.06,\ b_{xy} = 0.47$; (*ii*) $b_{yx} = 0.54,\ b_{xy} = 0.66$; (*iii*) $b_{yx} = 0.16,$ $b_{xy} = 0.04$; (*iv*) $b_{yx} = 0.64,\ b_{xy} = 1.5$; (*v*) $b_{yx} = 0.81,\ b_{xy} = 1.16$ }

9. (*a*) Explain why the regression line of Y on X is not necessarily the same as the regression line of X on Y. How would you decide which is the appropriate regression in any particular situation. Answer the following.?

(*i*) When do the two lines coincide?

(*ii*) When are they at right angles?

{ (*i*) Exact linear relation. (*ii*) Uncorrelated X, Y (*i. e.*, $\rho = 0$) }

(*b*) Calculate the coefficient of correlation and obtain the lines of regression from the following data

Price x_i	3	4	5	6	7	8	9	10	11	12
Demand y_i	25	24	20	20	19	17	16	13	10	6

($r = -0.98,\ \hat{y} = 31.45 - 1.93\,x,\ \hat{x} = 16 - 0.5\,y$)

(*c*) Given the following data:

$n = 100,\quad \Sigma x_i = 5000,\quad \Sigma y_i = 6000,$

$\Sigma x_i y_i = 300300,\quad \Sigma x_i^2 = 250400,\quad \Sigma y_i^2 = 360900,$

Calculate

(*i*) s_x, s_y and r,

(*ii*) regression lines,

(*iii*) estimate the value of y for $x = 55$.

{ (*i*) $s_x = 2,\ s_y = 3,\ r = 0.5$, (*ii*) $\hat{y} = 22.5 + 0.75\,x,\ \hat{x} = 30.2 + 0.33\,y,$ (*iii*) 63.75 }

10. (*a*) Given the following data:

$$n = 10, \quad \Sigma x_i = 120, \quad \Sigma y_i = 250, \quad \Sigma x_i\, y_i = 3070.7, \quad s_x = 3.5, \quad s_y = 7.2$$

Calculate regression lines.

($\hat{y} = 18.04 + 0.58\,x$; $\hat{x} = 8.50 + 0.14\,y$)

(*b*) Given that means and variances of two series X and Y are

	X-series	*Y*-series
Mean:	25	38
Variance:	25	36

The correlation coefficient between X and Y is 0.75. Estimate the most plausible value of Y for $x = 40$ and most plausible value of X for $y = 58$.

($\hat{y} = 15.5 + 0.9\,x$, 51.5; $\hat{x} = 1.25 + 0.625\,y$, 37.5)

(*c*) If the mean height of 500 fathers is 68.65 inches with standard deviation of 2.8 inches and the mean height of their youngest sons is 69.65 inches with standard deviation of 2.85 inches and the coefficient of correlation between them is 0.52 obtain the two equations of the lines of regression in the simplest form.

($\hat{y} = 33.27 + 0.53\,x$; $\hat{x} = 33.13 + 0.51\,y$)

11. (*a*) Given the following data:

$$\bar{x} = 54, \quad \bar{y} = 28, \quad b_{yx} = -1.5, \quad b_{xy} = -0.2$$

Show that the two estimated lines of regression intersect at the point $(\bar{x}, \bar{y})$. Estimate the value of X when $Y = 30$ and the value of Y when $X = 55$.

Hint: *Show that the estimated value of X for $Y = \bar{y} = 28$ is 54 and the estimated value of Y for $X = \bar{x} = 54$ is 28.*

($\hat{y} = 26.5$; $\hat{x} = 53.6$)

(b) For a given set of data, the least squares regression lines are

Estimated regression line of Y on X: $\hat{y} = 20.8 - 0.219\,x$

Estimated regression line of X on Y: $\hat{x} = 16.2 - 0.785\,y$

Find the product moment correlation coefficient for the data.

($r = -0.415$)

12. (*a*) For the following set of data, use $u_i = x_i - 1000$, and $v_i = (y_i - 250)/5$ to find the product moment correlation coefficient and the least squares lines of regression of Y on X and of X on Y.

x_i	1000	1012	1009	1007	1010	1015	1010	1011
y_i	235	240	245	250	255	260	265	270

(0.583; $\hat{y} = -1386.1346 + 1.6235\,x$; $\hat{x} = 956.326 + 0.2096\,y$)

(*b*) On each of 30 items, two measurements are made on the variables X and Y. The following summations are given

$$\Sigma x_i = 15, \quad \Sigma y_i = -6, \quad \Sigma x_i^2 = 61, \quad \Sigma y_i^2 = 90, \quad \Sigma x_i\, y_i = 56$$

Calculate the product moment correlation coefficient and obtain the lines of regression of Y on X and X on Y. If the variable X is replaced by U where $u_i = (x_i - 1)/2$, find the correlation coefficient between U and Y and the regression lines of Y on U and U on Y.
(0.856; $\hat{y} = -0.751 + 1.10x$; $\hat{x} = 0.633 + 0.664y$; 0.856; $\hat{y} = 0.351 + 2.21u$; $\hat{u} = -0.184 + 0.332y$)

(c) The following table shows the marks in statistics and mathematics obtained by 10 students from a large group of students.

Marks in Statistics	x_i	75	80	93	65	87	71	98	68	84	77
Marks in Mathematics	y_i	82	78	86	72	91	80	95	72	89	74

Estimate the linear regression function considering
(*i*) X as independent variable,
(*ii*) Y as independent variable.
($\hat{y} = 29.13 + 0.661x$; $\hat{x} = -14.39 + 1.15y$)

13. (a) A random sample of 20 pairs of observations (x_i, y_i) gave the following

$$\bar{x} = 2, \quad \bar{y} = 8, \quad \sum x_i^2 = 180, \quad \sum y_i^2 = 1424, \quad \sum x_i y_i = 404$$

Estimate the linear regression function taking (*i*) X as independent variable, (*ii*) Y as independent variable.
{ $\hat{y} = 6.32 + 0.84x$; $\hat{x} = -2.67 + 0.583y$ }

(b) Given the following data:

$$n = 5, \quad \sum x_i = 15, \quad \sum y_i = 25, \quad \sum (x_i - \bar{x})^2 = 10, \quad \sum (y_i - \bar{y})^2 = 26,$$

$\sum (x_i - \bar{x})(y_i - \bar{y}) = 13$. Determine the two regression lines.
($\hat{y} = 1.1 + 1.3x$; $\hat{x} = 0.5 + 0.5y$)

(c) The correlation coefficient between the two variables X and Y is $r = 0.60$. If $s_x = 1.50$, $s_y = 2.00$, $\bar{x} = 10$ and $\bar{y} = 20$, find the equations of the two regression lines of Y on X and X on Y.
($\hat{y} = 12 + 0.8x$; $\hat{x} = 1 + 0.45y$)

◆

Exercise 14.2
Objective Questions

1. Fill in the blanks.

(*i*) The ———— is a relationship that describes the dependence of the expected value of the dependent random variable for a given value of the independent non-random variable. (*regression*)

(*ii*) The variable, that forms the basis of estimation, is called ————. (*regressor*)

(*iii*) The variable, whose resulting value depends upon the selected value of the independent variable, is called ————. (*regressand*)

(*iv*) The ———— diagram is a set of points in a rectangular co-ordinate system, where each point represents an observed pair of values. (*scatter*)

(*v*) The principle of least squares is used for finding the ———— a and b of the parameters α and β. (*estimates*)

(*vi*) The ———— regression line always passes through $(\bar{x}, \bar{y})$. (*estimated*)

2. Mark the statements as true or false.

(*i*) The simple linear regression model contains two parameters α and β. (*false*)

(*ii*) The simple linear regression model contains four parameters α, β, $\mu_{Y|x}$ and σ^2. (*true*)

(*iii*) The simple linear regression model is simple in that there is only one independent variable. (*true*)

(*iv*) The parameter α is called the slope and the parameter β is the intercept of the regression line. (*false*)

(*v*) The regression coefficient is denoted by α. (*false*)

(*vi*) The parameter α is called the y-intercept of the regression line. (*true*)

(*vii*) In a regression analysis the independent variable is always prefixed while the dependent variable is random. (*true*)

(viii) The principle of least squares says that the sum of squares of the residuals of observed values from their corresponding estimated values should be the least possible. (*true*)

(*ix*) The principle of least squares is used for finding the estimates a and b of the parameters α and β. (*true*)

(*x*) The constant b estimates the parameter β representing the slope of the regression line. (*true*)

(*xi*) The regression coefficient b is independent of change in origin and scale. (*false*)

(*xii*) The estimated regression equation of Y on x is used to estimate the mean value of Y for a given value of x. (*true*)

3. Fill in the blanks.

(*i*) The correlation analysis is possible when both the variables X and Y are ————. (*random*)

(*ii*) If the two variables move in the ———— direction, the correlation is positive. (*same*)

(*iii*) If the two variables move in ———— directions, the correlation is negative. (*opposite*)

(*iv*) The correlation coefficient is ———— of the change in origin and unit of measurement. (*independent*)

(*v*) The correlation coefficient r is the ———— mean of the two regression coefficients. (*geometric*)

(*vi*) $r = 0$ indicates that the two variables are linearly ———— (*independent*)

4. Mark off the following statements true or false.

(*i*) The strength of covariability between two random variables is called correlation. (*true*)

(*ii*) The sample correlation coefficient R is a point estimator of the population correlation coefficient ρ. (*true*)

(*iii*) The correlation coefficient r is not symmetrical with respect to X and Y. (*false*)

(*iv*) The correlation coefficient changes with a change in origin. (*false*)

(*v*) The correlation coefficient is not affected by change in origin. (*true*)

(*vi*) The correlation coefficient is not independent of the origin and the unit of measurement. (*false*)

(*vii*) The correlation coefficient is a pure number which is unitless. (*true*)

(*xiii*) The correlation coefficient r always lies between -1 and 1. (*true*)

(*xiv*) $r = 1$ indicates perfect positive correlation between the two variables and slope is positive. (*true*)

(*xv*) $r = -1$ indicates perfect negative correlation between the two variables and slope is negative. (*true*)

5. Mark off the following statements true or false.

(*i*) $r = 0$ indicates that one regression line is horizontal and the other regression line is vertical. (*true*)

(*ii*) The correlation coefficient r is the geometric mean of the two regression coefficients. (*true*)

(*iii*) Each of the two estimated regression lines passes through the point $(\bar{x}, \bar{y})$. (*true*)

(*iv*) We can always estimate the X and Y values from the regression equation of Y on X. (*false*)

(*v*) The regression coefficient of X on Y is -1.2 and of Y on X is 0.3. (*false*)

(*vi*) The regression coefficient of X on Y is -1.2 and of Y on X is -0.3. (*true*)

(*vii*) The regression coefficient of X on Y, regression coefficient of Y on X and correlation coefficient have same sign. (*true*)

(*viii*) If the regression coefficient of X on Y is -1.2 and of Y on X is -0.3, the correlation coefficient is 0.6. (*false*)

(*ix*) The regression coefficient of X on Y is always equal to the regression coefficient of Y on X. (*false*)

15 ASSOCIATION

Many experiments, particularly in social sciences, result in observations that are only classified into categories so that the data can consist of frequency count for the categories. For example, the classification of people into income groups as very rich, moderate, or poor; manufactured items may be classified as being excellent, good, poor, or scrap condition; in a survey of job compatibility employed persons may be classified as being satisfied, neutral, or dissatisfied with their jobs; in plant breeding, the offsprings of a cross fertilization may be grouped into several genotypes; rainfall may be classified heavy, moderate, or light; each household may be classified as owning no cars, one car, or two or more cars. Our aim here is to present some inferential procedures that can be used to study data that are classified into multiple categories.

15.1 MULTINOMIAL POPULATIONS

When each element of a population is assigned to one and only one of more than two attribute categories, the population is called a *multinomial population*.

15.2 ATTRIBUTE (*QUALITATIVE VARIABLE*)

A characteristic which varies only in quality from one individual to another, is called an *attribute*. Examples of attributes are: marital status, education level, blindness, smoking, richness, beauty *etc*. It is not possible to measure an attribute quantitatively. The quantitative data relating to an attribute may be obtained simply by noting its presence or absence in the objects, and then counting that how many do or do not possess that attribute.

15.2.1 Class and Class Frequency. A *class* is a set of the objects which are sharing a given characteristic. A *class frequency* is the number of observations (or objects) which are distributed in a class.

15.2.2 Classification of Objects. The objects (or individuals) can be divided into two distinct, mutually exclusive and complementary classes according to whether the objects do or do not possess a particular attribute. This process of dividing the objects into two mutually exclusive classes is called *dichotomy*.

If several attributes are noted, the process of classification may be continued indefinitely. The objects that are classified according to as they do or do not possess the first attribute can further be subdivided according to as they do or do not possess the second attribute and the objects of each of these subclasses can still further be subdivided according to as they do or do not possess the third attribute, and so on, every class being divided into two subclasses at each step. For example, the members of the population of district Lahore may be classified according to sex as males or females; the members of each sex may be further subdivided according to marital status as married or unmarried; that results into the married males, unmarried males, married females or unmarried females; the members of these four classes may be still further subdivided according to educational status as literate or illiterate.

15.2.3 Notations and Terminology. For theoretical study it is necessary to have some notations to represent different classes and their class frequencies. The capital Latin letters A, B, $\cdots$ are used to denote the attributes and their presence. The Greek letters α, β, $\cdots$ are used to denote the absence of these attributes. Thus A will denote that the object possesses the attribute A and α will denote that the object does not possess the attribute A; B will denote that the object possesses the attribute B, and β will denote that the object does not possess the attribute B. Hence "α" means "not A", "β" means "not B".

Class frequencies will be denoted by enclosing the class by symbols in brackets. Thus (A) denotes the number of objects possessing the attribute A; (αB) denotes the number of objects possessing the attribute B but not the attribute A.

The attributes denoted by $A, B, \cdots$ are called *positive attributes* and their contraries denoted by $\alpha, \beta, \cdots$ are called *negative attributes*. Thus the classes A, B and AB represented by positive attributes are called *positive classes*; the classes α, β and $\alpha\beta$ represented by negative attributes are called *negative classes*; and the classes $A\beta$, αB, *etc.* represented by both positive as well as negative attributes are called *contrary classes*.

15.2.4 Order of Classes. *Order of class* is known by the number of attributes specifying the class, *e. g.*, a class specified by one attribute is known as the class of order 1, the classes specified by two attributes are called as the classes of order 2; and the classes specified by three attributes are known as the classes of order 3. The total number of observations denoted by n is called the frequency of the class of order zero since no attributes are specified.

In the study of only one attribute A, we have the following frequencies

Frequency of the class of order zero : n

Frequencies of the classes of order 1 : (A), (α)

In the study of two attributes A and B, we have the following frequencies

Frequency of the class of order zero : n

Frequencies of the classes of order 1 : (A), (α), (B), (β)

Frequencies of the classes of order 2 : (AB), $(A\beta)$, (αB), $(\alpha\beta)$

These observed frequencies can be expressed in the form of a 2×2 table as

Attribute A	Attribute B		Total
	B	β	
A	(AB)	$(A\beta)$	(A)
α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	n

15.2.5 Number of Class Frequencies. If we include the total number of observations n as a frequency of the class of order zero, then in general, for k attributes the total number of class frequencies would be $(3)^k$. Thus in case of only one attribute the total number of class frequencies would be $(3)^1 = 3$; for two attributes it is $(3)^2 = 9$, and so on.

15.2.6 Ultimate Class Frequency. The frequencies of classes of the highest order are called *ultimate class frequencies.* The number of ultimate class frequencies for k attributes is given by $(2)^k$. Thus in case of two attributes the number of ultimate classes is $(2)^2 = 4$, and so on.

If n is included as a positive class, then for k attributes the number of positive classes is the same as the number of ultimate classes. For two attributes, the positive classes are n, (A), (B), (AB) and the ultimate classes are (AB), $(A\beta)$, (αB), $(\alpha\beta)$.

It is interesting to note a very simple result that any class frequency can always be expressed in terms of the class frequencies of higher order. Any class can always be expressed as a sum of its two subclasses produced by dichotomizing it for the study of a new characteristic. For example, in the study of two attributes, we may have:

$$n = (A) + (\alpha) \qquad n = (B) + (\beta)$$
$$(A) = (AB) + (A\beta) \qquad (B) = (AB) + (\alpha B)$$
$$(\alpha) = (\alpha B) + (\alpha\beta) \qquad (\beta) = (A\beta) + (\alpha\beta)$$

15.2.7 Consistence of data. The class frequencies that have been observed in one and the same population are said to be *consistent*, if they conform with one another and do not conflict each other. In the study of attributes, no class frequency can ever be negative. If any class frequency is negative the data are said to be inconsistent. Inconsistency may be due to wrong counting, inaccurate additions or subtractions or due to misprints. The necessary and sufficient condition for the consistence of a set of class frequencies is that no ultimate class frequency should be negative. To test the consistence of data, we calculate the ultimate class frequencies from the given data and if any of the ultimate class frequencies turns out to be negative, data will be *inconsistent*. If no ultimate class frequency is negative, the data are consistent. It is however important to note that the consistence of data is no proof of accurate count, accurate additions or subtractions or the absence of misprints.

15.3 INDEPENDENCE OF ATTRIBUTES

If in a sample of size n, the class frequencies of attributes A, B and AB are represented by (A), (B) and (AB). Then we have

$$\text{Proportion of individuals possessing } A = \frac{(A)}{n}$$

$$\text{Proportion of individuals possessing } B = \frac{(B)}{n}$$

$$\text{Proportion of individuals possessing } AB = \frac{(AB)}{n}$$

The two attributes A and B are said to be independent if,

$$\text{Proportion of } AB = (\text{Proportion of } A)(\text{Proportion of } B)$$

$$\frac{(AB)}{n} = \frac{(A)}{n} \times \frac{(B)}{n}$$

$$(AB) = \frac{(A)(B)}{n}$$

In case of independence of attributes A and B, the 2×2 table must have the form

Attribute A	Attribute B		Total
	B	β	
A	$\frac{(A)(B)}{n}$	$\frac{(A)(\beta)}{n}$	(A)
α	$\frac{(\alpha)(B)}{n}$	$\frac{(\alpha)(\beta)}{n}$	(α)
Total	(B)	(β)	n

***Example* 15.1** *If there are* 144 *A's and* 384 *B's in* 1024 *observations. How many* (*i*) *AB's and* (*ii*) $\alpha\beta$'s *will there be for A and B* being independent.

Solution. We have $n = 1024, \quad (A) = 144, \quad (B) = 384$

For A and B being independent, we must have

(*i*) $$(AB) = \frac{(A)(B)}{n} = \frac{(144)(384)}{1024} = 54$$

(*ii*) $$(\alpha) = n - (A) = 1024 - 144 = 880$$

$$(\beta) = n - (B) = 1024 - 384 = 640$$

$$(\alpha\beta) = \frac{(\alpha)(\beta)}{n} = \frac{(880)(640)}{1024} = 550$$

***Example* 15.2** *If the A's are* 60%, *the B's are* 40%, *of the whole number of observations, what must be the percentage of AB's in order that we may conclude that A and B are independent*?

Solution. Let $n = 100$, then $(A) = 60, \quad (B) = 40$

For A and B being independent, we must have

$$(AB) = \frac{(A)(B)}{n} = \frac{60 \times 40}{100} = 24$$

There must be 24% *AB*'s to justify the conclusion that A and B are independent.

***Example* 15.3** *Given the following data. Find whether A and B are independent or associated.*

(*i*) $n = 150, \quad (A) = 30, \quad (B) = 60, \quad (AB) = 12$

(*ii*) $(AB) = 256, \quad (\alpha\beta) = 144, \quad (A\beta) = 48, \quad (\alpha B) = 768$

Solution.

(*i*) Observed frequency of *AB*'s $= (AB) = 12$

Expected frequency of *AB*'s $= \frac{(A)(B)}{n} = \frac{(30)(60)}{150} = 12$

Since $(AB) = \frac{(A)(B)}{n}$, the attributes A and B are independent.

(*ii*) We have the 2×2 table as

Attribute A	Attribute B		Total
	B	β	
A	$(AB) = 256$	$(A\beta) = 48$	$(A) = 304$
α	$(\alpha B) = 768$	$(\alpha\beta) = 144$	$(\alpha) = 912$
Total	$(B) = 1024$	$(\beta) = 192$	$n = 1216$

Observed frequency of AB's $= (AB) = 256$

Expected frequency of AB's $= \dfrac{(A)(B)}{n} = \dfrac{(304)(1024)}{1216} = 256$

Since $(AB) = \dfrac{(A)(B)}{n}$, the attributes A and B are independent.

15.4 ASSOCIATION OF ATTRIBUTES (*CORRELATION OF QUALITATIVE VARIABLES*)

The two attributes A and B are said to be *associated* if they are not independent, *i. e.*,

$$(AB) \neq \frac{(A)(B)}{n}$$

Association of attributes may be classified as positive or negative.

15.4.1 Positive Association. The two attributes A and B are *positively associated* or *simply associated*, if

$$(AB) > \frac{(A)(B)}{n}$$

15.4.2 Negative Association. The two attributes A and B are *negatively associated* or *simply disassociated*, if

$$(AB) < \frac{(A)(B)}{n}$$

It should be noted that disassociation does not imply independence.

15.4.3 Complete Association and Disassociation. There will be complete (or perfect positive) association between two attributes A and B if one of them cannot occur without the other, though the other may occur without the one, that is, if

(*i*) $(A) = (B) \Rightarrow$ all A'*s* are B'*s* and all B'*s* are A'*s*

(*ii*) $(A) < (B) \Rightarrow$ all A'*s* are B'*s*

(*iii*) $(B) < (A) \Rightarrow$ all B'*s* are A'*s*

There will be complete disassociation (or perfect negative association) between two attributes A and B if none of A'*s* is B'*s* and none of α'*s* is β'*s*.

15.4.4 Coefficient of Association. The strength of association, between two attributes A and B, is known as *coefficient of association.*

The Yule's coefficient of association, denoted by Q, is defined as :

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

This coefficient lies between -1 and $+1$.

If $Q = 0$, the two attributes are independent.

If $Q = 1$, the two attributes are completely associated.

If $Q = -1$, the two attributes are completely disassociated.

***Example* 15.4** *Given the following* :

$$(AB) = 110, \quad (\alpha B) = 90, \quad (A\beta) = 290, \quad (\alpha\beta) = 510$$

Discuss association.

Solution. We have the 2×2 table as

Attribute A	Attribute B		Total
	B	β	
A	$(AB) = 110$	$(A\beta) = 290$	$(A) = 400$
α	$(\alpha B) = 90$	$(\alpha\beta) = 510$	$(\alpha) = 600$
Total	$(B) = 200$	$(\beta) = 800$	$n = 1000$

Observed frequency of AB's $= (AB) = 110$

Expected frequency of AB's $= \frac{(A)(B)}{n} = \frac{(400)(200)}{1000} = 80$

Since $(AB) > \frac{(A)(B)}{n}$, the attributes A and B are positively associated.

***Example* 15.5** 1660 *candidates appeared for a competitive examination and* 422 *were successful.* 256 *had attended a coaching class and of these* 150 *came out successful. Find the coefficient of association between success and coaching a class.*

Solution. Let A represent success and B represent attending coaching class, then we have

$$n = 1660, \quad (A) = 422, \quad (B) = 256, \quad (AB) = 150$$

Attribute A	Attribute B		Total
	B	β	
A	$(AB) = 150$	$(A\beta) = 422 - 150 = 272$	$(A) = 422$
α	$(\alpha B) = 256 - 150 = 106$	$(\alpha\beta) = 1404 - 272 = 1132$	$(\alpha) = 1660 - 422 = 1238$
Total	$(B) = 256$	$(\beta) = 1660 - 256 = 1404$	$n = 1660$

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{(150)(1132) - (272)(106)}{(150)(1132) + (272)(106)} = 0.71$$

Exercise 15.1

1. (*a*) Distinguish between attribute and variable. Define positive classes; negative classes and ultimate classes.

(*b*) Given the following ultimate class frequencies, find the frequencies of the positive and negative classes and the whole number of observations n.

$(AB) = 95, \quad (A\beta) = 55, \quad (\alpha B) = 85, \quad (\alpha\beta) = 45$

$\{ n = 280, \ (A) = 150, \ (B) = 180, \ (\alpha) = 130, \ (\beta) = 100 \}$

2. (*a*) Given the following frequencies of the positive classes, find the frequencies of ultimate classes.

$n = 250, \quad (A) = 80, \quad (B) = 100, \quad (AB) = 70.$

$\{ (A\beta) = 10, \quad (\alpha B) = 30, \quad (\alpha\beta) = 140, \quad (AB) = 70 \}$

(*b*) Measurements are made on a thousand husbands and a thousand wives. If the measurements of husbands exceed the measurements of the wives in 800 cases for one measurements, in 700 cases for another, and in 660 cases for both measurements, in how many cases will both the measurements on the wife exceed the measurement on the husband?

(160)

(*c*) Given that $(A) = (\alpha) = (B) = (\beta) = n/2$, show that

(*i*) $(AB) = (\alpha\beta)$ (*ii*) $(A\beta) = (\alpha B)$

3. (*a*) Define the consistence of the data.

(*b*) Find whether the data given below in each case are consistent?

(*i*) $n = 120, \quad (A) = 82, \quad (AB) = 90$

(*ii*) $n = 50, \quad (A) = 40, \quad (B) = 32, \quad (AB) = 15$

(*iii*) $n = 1000, \quad (AB) = 200, \quad (A\beta) = 350, \quad (\alpha B) = 500$

{ (*i*) Not consistent since $(A\beta) = -8$, (*ii*) Not consistent since $(\alpha\beta) = -7$, (*iii*) Not consistent since $(\alpha\beta) = -50$ }

(*c*) Comment on the following data contained in a report: 100 students appeared in a test of whom 80 passed in Statistics: 70 passed in Mathematics and 48 passed in both the subjects.

{ Not consistent, since $(\alpha\beta) = -2$ }

4. (*a*) What is meant by independence of attributes.

(*b*) There is 240 *A*'s and 270 *B*'*s* in 600 observations. What would be the number of *AB* if *A* and *B* are independent.

$\{ (AB) = 108 \}$

(*c*) If *A*'s are 60% and *B*'s are 40% of the whole number of observations, what must be the percentage of *AB*'s in order that we conclude that *A* and *B* are independent.

{ *AB*'s are 24% }

5. (*a*) When are two attributes independent, positively associated, negatively associated?

(*b*) Given the following data, determine the nature of association between the attributes *A* and *B*, *i. e.*, find whether *A* and *B* are independent, positively associated or negatively associated.

(*i*) $(A) = 30, \quad (B) = 60, \quad (AB) = 12, \quad n = 150$

(*ii*) $(AB) = 110, \quad (\alpha B) = 90, \quad (A\beta) = 290, \quad (\alpha\beta) = 510$

(*iii*) $(A) = 415$, $(AB) = 147$, $(\alpha) = 285$, $(\alpha\beta) = 170$

{ (*i*) Independent, (*ii*) Positively associated, (*iii*) Negative associated }

6. (*a*) What is meant by association of attributes?

(*b*) Explain the difference between the following with examples.

(*i*) Attribute and variable,

(*ii*) Correlation and association,

(*iii*) Positive association and negative association

7. (*a*) Find the association between injection against typhoid and exemption from attack from the following contingency table

Attribute	Attacked	Not attacked
Inoculated	528	25
Not inoculated	790	175

($Q = 0.65$)

(*b*) Calculate the coefficient of the association between the intelligence of fathers and sons in the following data:

Intelligent fathers with intelligent sons = 265

Intelligent fathers with dull sons = 100

Dull fathers with intelligent sons = 95

Dull fathers with dull sons = 450

($Q = 0.85$)

(*c*) Find if there is any association between the tempers of bothers and sisters from the following data :

Good natured bothers and good natured sisters = 1230

Good natured bothers and sullen sisters = 850

Sullen bothers and good natured sisters = 530

Sullen bothers and sullen sisters = 980

($Q = 0.46$)

8. (*a*) 750 students appeared in an examination and 470 were successful. 465 had attended classes and 58 of them failed. Calculate the coefficient of association to discuss association between attending classes and success.

($Q = 0.92$, highly positive association)

(*b*) 100 students appeared in an examination, and 50 failed in Mathematics, 60 failed in Statistics and 40 failed in both. Find if there is any association between the failing in Mathematics and Statistics.

($Q = 0.71$)

(*c*) Can vaccination be regarded as preventive measure for small pox from the following data: "Of 1482 persons in a locality exposed to small pox, 368 in all were attacked. Of 1482 persons, 343 persons, had been vaccinated and of these 35 were attacked".

($Q = -0.57$)

◆

15.5 TWO DIMENSIONAL COUNT DATA: CONTINGENCY TABLE

A simple random sample of n elements selected from a bivariate multinomial population that has been classified into r categories $A_1, A_2, \cdots, A_r$ of attribute A and c categories $B_1, B_2, \cdots, B_c$ of attribute B will produce a two-way frequency table which is called an $r \times c$ *contingency table* — a name due to Karl Pearson. A contingency table is made up of the observed frequencies relative to the two attributes and their categories which is generally presented in the following tabular form, with rows representing the r categories $A_1, A_2, \cdots, A_r$ of attribute A and columns representing c categories $B_1, B_2, \cdots, B_c$ of attribute B.

15.5.1 Cell Frequency. The number of observations falling in a particular cell is called the *cell frequency*.

An $r \times c$ Contingency Table

Attribute A	Attribute B					
	B_1	B_2	B_j	$\cdots$	B_c	Row total
A_1	o_{11}	o_{12}	o_{1j}	$\cdots$	o_{1c}	$o_{1\cdot}$
A_2	o_{21}	o_{22}	o_{2j}	$\cdots$	o_{2c}	$o_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_i	o_{i1}	o_{i2}	o_{ij}	$\cdots$	o_{ic}	$o_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	o_{r1}	o_{r2}	o_{rj}	$\cdots$	o_{rc}	$o_{r\cdot}$
Column total	$o_{\cdot 1}$	$o_{\cdot 2}$	$o_{\cdot j}$	$\cdots$	$o_{\cdot c}$	n

The table shows, in all, $k = rc$ cells or categories. The symbol O_{ij} denotes the number of sample observations in the (i, j) category of attributes A and B, respectively, for $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$. The entries in the table represent the realizations o_{ij} the observed frequencies of the random variables O_{ij}. Note that the i-th row total is the observed frequency of the i-th category of attribute A summed over all categories of attribute B. Similarly, the j-th column total is the observed frequency of the j-th category of attribute B summed over all categories of attribute A. Let

$$o_{i\cdot} = \sum_{j=1}^{c} o_{ij} \qquad \text{for } i = 1, 2, \cdots, r$$

$$o_{\cdot j} = \sum_{i=1}^{r} o_{ij} \qquad \text{for } j = 1, 2, \cdots, c$$

denote the row and column sums, respectively, where the "dot" notation indicates the subscript over which summation has taken place. That is

o_{ij} = Observed frequency of $A_i \cap B_j$

$o_{i\cdot}$ = Observed frequency of A_i, *i. e.*, i-th row total

$o_{\cdot j}$ = Observed frequency of B_j, *i. e.*, j-th column total

$$\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} = \sum_{i=1}^{r} o_{i\cdot} = \sum_{j=1}^{c} o_{\cdot j} = n$$

15.6 TEST FOR STATISTICAL INDEPENDENCE

In analyzing bivariate multinomial populations, the first-step of a typical inferential aspect of interest usually is whether the two attributes are statistically independent or whether certain levels of one attribute tend to be associated or contingent with some levels of another attribute. If they are independent, we know that there is no relationship between them. If it turns out that they are not independent and a relationship does exit between the two attributes, the next step in the analysis then is to study the nature of the relationship. We begin with the first step of the analysis, testing whether or not the two attributes are independent.

We are concerned with testing the null hypothesis that the two criteria of classification are independent. Recall, if two classifications are independent of each other, a cell probability will equal the product of its respective row and column probabilities in accordance with multiplicative law of probability. Therefore, the null hypothesis stating that the events $A_1, A_2, \cdots, A_r$ are independent of events $B_1, B_2, \cdots, B_c$ can be rephrased

$$P(A_i \cap B_j) = P(A_i)\,P(B_j) \qquad \text{for all } i = 1, 2, \cdots, r \text{ and } j = 1, 2, \cdots, c.$$

Thus the null and alternative hypotheses for a test of statistical independence are

Null hypothesis H_0: A_i and B_j are independent for all cells (i, j)

Alternative hypothesis H_1: A_i and B_j are not independent for some (i, j)

Here, H_0 represents statistical independence and H_1 represent statistical dependence.

The problem now becomes testing the goodness of fit for the model of independence. We compare the observed frequencies o_{ij} with the expected frequencies $E(O_{ij})$ that are expected if the attributes are independent. Under the null hypothesis of independence of attributes the estimate of expected frequency $E(O_{ij})$ is

$$e_{ij} = \frac{o_{i\cdot}\,o_{\cdot j}}{n} = \frac{(i\text{-th row total})\,(j\text{-th column total})}{\text{number of observations}}$$

The test statistic then becomes

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

which has an approximate chi-square distribution with $\nu = (r - 1)(c - 1)$ degrees of freedom for large n.

In this case, we base the test statistic on the expected number of elements e_{ij} in the sample from each category if H_0 is true and the postulated proportions hold. The farther the observed frequency o_{ij} departs in either direction from the expected frequency e_{ij}, the large is $(o_{ij} - e_{ij})^2$ and hence the larger is χ^2. On the other hand, if there is perfect agreement between the observed and expected frequencies, *i. e.*, o_{ij} and e_{ij} are identical for all classes, $\chi^2 = 0$ because each $(o_{ij} - e_{ij})^2 = 0$. If all the observed frequencies o_{ij} are close to the expected frequencies e_{ij} supporting H_0, the value of χ^2 will be near to zero; if o_{ij} are far from e_{ij}, indicating rejection of H_0, the χ^2 will assume a large positive value. It follows, therefore, that for a given level of significance α the critical region is the upper tail of chi-square distribution with $\nu = (r - 1)(c - 1)$ degrees of freedom, *i. e.*,

Critical region: $\chi^2 > \chi^2_{\nu;1-\alpha}$

On the question how large a sample size should be, we know that this test is based on the normal approximation to the binomial, a fairly conservative rule of thumb is that the approximation is adequate if each $e_{ij} \geq 5$. If there are not at least 5 items, the value of chi-square is inflated because squared differences are divided by a very small size expected frequency in $\chi^2 = \sum\{(o_{ij} - e_{ij})^2 / e_{ij}\}$. However, if the cells have too small expected frequencies the condition of at least 5 items in each expected frequency class can also be accomplished by combining neighbouring row or column class, but for pair of rows or columns that is combined the number of rows or columns for degrees of freedom is reduced by one.

15.6.1 Assumptions. To conduct a valid test of hypothesis for independence using data from a contingency table, the following conditions must be met.

(*i*) A simple random sample of size n has been selected from a bivariate multinomial population.

(*ii*) The sample size n is reasonably large so that for each cell, the estimated expected frequency must be at least 5.

15.6.2 Yates' Correction. To improve the approximation to the χ^2 distribution and thus be able to obtain a more exact probability value from the χ^2 table, F. Yates has proposed a *correction for continuity*, applicable when the criterion has a single degree of freedom. The correction is intended to make the actual distribution of the criterion, as calculated from discrete data, more nearly like the χ^2 distribution based on normal deviations. The relation $Z^2 = \chi^2$ between Z and χ^2 holds only for a single degree of freedom. The approximation calls for the absolute value of each deviation to be decreased by $1/2$, because for two celled tables, the deviations are always equal in magnitude but opposite in sign. Therefore

$$\text{Adjusted } \chi^2 = \sum \frac{(|o - e| - 0.5)^2}{e}$$

Thus Yates' correction is analogous to the continuity correction which is applied in the normal approximation to the binomial distribution. There is a tendency to under estimate the

probability, which means that the probability of rejecting the hypothesis will be increased. Adjustment results in a lower chi-square. Consequently, in testing hypothesis, it is worthwhile only when unadjusted χ^2 is greater than tabulated χ^2 at the desired probability level. When n (or e) is large continuity correction has little effect, but when e's are small, it should be applied. When $|o - e|$ is less than 0.5, the continuity correction should be omitted.

15.6.3 Coefficient of Contingency. The *coefficient of contingency* is a measure of the strength of association on a numerical scale as an index of association between two criteria of classification. When the test for statistical independence leads to the conclusion of dependence, we may wish to measure the strength of association between two criteria of classification. Insofar as the χ^2 statistic represents an over all deviation from the model of independence, it is intuitively reasonable to use this statistic to gauge the strength of this association. We may call χ^2 as the "square contingency". But in applying the χ^2 statistic as a measure of association the limitation is that the number of degrees of freedom attached to this statistic depends upon the dimensionality of the contingency table. A χ^2 value of 16.5 in a 2×2 contingency table would reflect a significant association, but this would not be so in a 6×8 contingency table. Several measures of association have been proposed to adjust the χ^2 statistic to a common scale that is irrespective of the dimensionality of the contingency table. We then write

$$\phi^2 = \frac{\chi^2}{n}$$

and call ϕ^2 as the "*mean square contingency*". In the following are two commonly used formulas, large values of a measure indicate a strong association and small values of a measure indicate a weak association between the two criteria of classification.

Pearson's coefficient of mean square contingency:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}, \qquad 0 \leq C \leq \sqrt{\frac{q-1}{q}}$$

where q represents the number of rows or columns, whichever is smaller, and n indicates the sample size.

***Example* 15.6** *Four hundred and ninety two candidates for scientific posts gave particulars of their university degrees and their hobbies. The degrees were in either mathematics, chemistry or physics and the hobbies could be classified roughly as music, craftwork, reading or drama. The data are presented concisely in the following contingency table.*

	Degree		
Hobby	*Mathematics*	*Chemistry*	*Physics*
Music	24	83	17
Craftwork	11	62	28
Reading	32	121	34
Drama	10	26	44

Discuss the association between the two criteria of classification, i.e., the degrees and hobbies. If the null hypothesis of independence is rejected, calculate the Pearson's coefficient of mean square contingency. What could be its maximum value for this contingency table.

Solution. The elements of the one-sided right tail test of hypothesis are

Null hypothesis H_0: The degree and hobby are independent.

Alternative hypothesis H_1: The degree and hobby are not independent.

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ follows an approximate chi-square distribution under H_0 with

Degrees of freedom: $\nu = (r - 1)(c - 1) = (4 - 1)(3 - 1) = 6,$

Critical value: $\chi^2_{\nu;1-\alpha} = \chi^2_{6;0.95} = 12.59$ (From Table 11)

Critical region: $\chi^2 > 12.59$

Decision rule: Reject H_0 if $\chi^2 > 12.59$, otherwise do not reject H_0.

Observed value: The observed frequencies o_{ij} are :

Hobby	Degree Mathematics: B_1	Chemistry: B_2	Physics: B_3	Row total
Music: A_1	24	83	17	$o_{1\cdot} = 124$
Craftwork: A_2	11	62	28	$o_{2\cdot} = 101$
Reading: A_3	32	121	34	$o_{3\cdot} = 187$.
Drama: A_4	10	26	44	$o_{4\cdot} = 80$
Column total	$o_{\cdot 1} = 77$	$o_{\cdot 2} = 292$	$o_{\cdot 3} = 123$	$n = 492$

The expected frequencies e_{ij} under the null hypothesis of independence are

$$e_{ij} = \frac{o_{i\cdot}\, o_{\cdot j}}{n} = \frac{(i\text{-th row total})\,(j\text{-th column total})}{\text{number of observations}}$$

which are given in the following table. Only $(r - 1)(c - 1) = (4 - 1)(3 - 1) = 6$ expected frequencies are obtained through this procedure. We could work through this procedure to give the other expected frequencies, but this is unnecessary, as the remaining frequencies can be found by using the fact that the sub-totals and totals must agree with those in observed data:

Hobby	Degree Mathematics: B_1	Chemistry: B_2	Physics: B_3	Row total
Music: A_1	$\frac{(124)(77)}{492} = 19.4$	$\frac{(124)(292)}{492} = 73.6$	31.0	124
Craftwork: A_2	$\frac{(101)(77)}{492} = 15.8$	$\frac{(101)(292)}{492} = 59.9$	25.3	101
Reading: A_3	$\frac{(187)(77)}{492} = 29.3$	$\frac{(187)(292)}{492} = 111.0$	46.7	187
Drama: A_4	12.5	47.5	20.0	80
Column total	77	292	123	492

The χ^2 -statistic is calculated as under

Cell (i,j)	Observed frequency o_{ij}	Expected frequency e_{ij}	$\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$
A_1B_1	24	19.4	1.09
A_1B_2	83	73.6	1.20
A_1B_3	17	31.0	6.32
A_2B_1	11	15.8	1.46
A_2B_2	62	59.9	0.07
A_2B_3	28	25.3	0.29
A_3B_1	32	29.3	0.25
A_3B_2	121	111.0	0.90
A_3B_3	34	46.7	3.45
A_4B_1	10	12.5	0.50
A_4B_2	26	47.5	9.73
A_4B_3	44	20.0	28.80
Total	492	492	$\chi^2 = 54.06$

Conclusion: Since $\chi^2 = 54.06 > 12.59$, we reject H_0 and conclude that the two criteria of classification are association.

Pearson's coefficient of mean square contingency:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{54.06}{492+54.06}} = 0.315$$

Maximum value of C for this contingency table:

$$\sqrt{\frac{q-1}{q}} = \sqrt{\frac{3-1}{3}} = 0.8165$$

***Example* 15.7** *Discuss the resemblances of stature of parents and off-springs for the following data*

	Parents			
Off-springs	*Very tall*	*Tall*	*Medium*	*Short*
Very tall	20	30	20	2
Tall	14	125	85	12
Medium	3	140	165	125
Short	3	37	68	151

Solution. The elements of the one-sided right tail test of hypothesis are

Null hypothesis: H_0: The stature of off-springs is independent of the stature of parents.

Alternative hypothesis: H_1: The stature of off-springs is not independent of the stature of parents.

Level of significance: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Test statistic: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ follows an approximate chi-square distribution under H_0 with

Degrees of freedom: $\nu = (r-1)(c-1) = (4-1)(3-1) = 6$, since first two column are pooled because $e_{11} = 2.88 < 5$.

Critical value: $\chi^2_{\nu;1-\alpha} = \chi^2_{6;0.95} = 12.59$ (From Table 11)

Critical region: $\chi^2 > 12.59$

Decision rule: Reject H_0 if $\chi^2 > 12.59$, otherwise do not reject H_0.

Observed value: The observed frequencies o_{ij} are given in the following contingency table:

Off-springs	Parents: Very tall: B_1	Tall: B_2	Medium: B_3	Short: B_4	Row total
Very tall: A_1	20	30	20	2	$o_{1.} = 72$
Tall: A_2	14	125	85	12	$o_{2.} = 236$
Medium: A_3	3	140	165	125	$o_{3.} = 433$
Short: A_4	3	37	68	151	$o_{4.} = 259$
Column total	$o_{.1} = 40$	$o_{.2} = 332$	$o_{.3} = 338$	$o_{.4} = 290$	$n = 1000$

The expected frequencies e_{ij} under the null hypothesis of independence are

$$e_{ij} = \frac{o_{i.}\, o_{.j}}{n} = \frac{(i\text{-th row total})\,(j\text{-th column total})}{\text{number of observations}}$$

which are given in the following table. Only $(r-1)(c-1) = (4-1)(4-1) = 9$ expected frequencies are obtained through this procedure. We could work through this procedure to give the other expected frequencies, but this is unnecessary, as the remaining frequencies can be found by using the fact that the sub-totals and totals must agree with those in observed data.

Off-springs	Parents: Very tall: B_1	Tall: B_2	Medium: B_3	Short B_4	Row total
Very tall: A_1	$\frac{(72)(40)}{1000}$ = 2.88	$\frac{(72)(332)}{1000}$ = 23.90	$\frac{(72)(338)}{1000}$ = 24.34	20.88	72
Tall: A_2	$\frac{(236)(40)}{1000}$ = 9.44	$\frac{(236)(332)}{1000}$ = 78.35	$\frac{(236)(338)}{1000}$ = 79.77	68.44	236
Medium: A_3	$\frac{(433)(40)}{1000}$ = 17.32	$\frac{(433)(332)}{1000}$ = 143.76	$\frac{(433)(338)}{1000}$ = 146.35	125.57	433
Short: A_4	10.36	85.99	87.54	75.11	259
Column total	40	332	338	290	1000

Combining the first and second columns of the expected frequencies, we get

Off spring	Parents Tall: B_1	Medium: B_2	Short: B_3
Very tall: A_1	2.88 + 23.90 = 26.78	24.34	20.88
Tall: A_2	9.44 + 78.35 = 87.79	79.77	68.44
Medium: A_3	17.32 + 143.76 = 161.08	146.35	125.57
Short: A_4	10.36 + 85.99 = 96.35	87.54	75.11

Accordingly, we combined the observed frequencies as under

Off spring	Parents Tall: B_1	Medium: B_2	Short: B_3
Very tall: A_1	20 + 30 = 50	20	2
Tall: A_2	14 + 125 = 139	85	12
Medium: A_3	3 + 140 = 143	165	125
Short: A_4	3 + 37 = 40	68	151

The χ^2 statistic is calculated as under

Cell (i, j)	Observed frequency o_{ij}	Expected frequency e_{ij}	$\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
$A_1 B_1$	50	26.78	20.13
$A_1 B_2$	20	24.34	0.77
$A_1 B_3$	2	20.88	17.07
$A_2 B_1$	139	87.79	29.87
$A_2 B_2$	85	79.77	0.34
$A_2 B_3$	12	68.44	46.54
$A_3 B_1$	143	161.08	2.03
$A_3 B_2$	165	146.35	2.38
$A_3 B_3$	125	125.57	0.00
$A_4 B_1$	40	96.35	32.96
$A_4 B_2$	68	87.54	4.36
$A_4 B_3$	151	75.11	76.68
Total	1000	1000	$\chi^2 = 233.13$

Conclusion: Since $\chi^2 = 233.13 > 12.59$, we reject H_0 and conclude that the Stature of off-springs is not independent of the stature of the parents.

Pearson's coefficient of mean square contingency :

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{233.13}{1000 + 233.13}} = 0.435$$

***Example* 15.8** *A random sample of* 30 *adults is classified according to the sex and the number of hours they watch television during a week :*

Time watching television	*Sex*	
	Male	*Female*
Over 25 *hours*	5	8
Under 25 *hours*	10	7

Using $\alpha = 0.01$ *test the hypothesis that a person's sex and time watching television are independent.*

Solution. The elements of the one-sided right tail test of hypothesis are

Null hypothesis H_0: The sex and time watching television are independent.

Alternative hypothesis H_1: The sex and time watching television are not independent.

Level of significance: $\alpha = 0.01 \quad \Rightarrow \quad 1 - \alpha = 0.99$

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$ follows an approximate chi-square distribution under H_0 with

Degrees of freedom: $\nu = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

Critical value: $\chi^2_{\nu;\,1-\alpha} = \chi^2_{1;\,0.99} = 6.63$

Critical region: $\chi^2 > 6.63$

Decision rule: Reject H_0 if $\chi^2 > 6.63$, otherwise do not reject H_0.

Observed value: The observed frequencies o_{ij} are given in the following table:

Time watching television	Sex		Row total
	Male: B_1	Female: B_2	
Over 25 hours: A_1	5	8	$o_{1.} = 13$
Under 25 hours: A_2	10	7	$o_{2.} = 17$
Column total	$o_{.1} = 15$	$o_{.2} = 15$	$n = 30$

The expected frequencies e_{ij} under the null hypothesis are

$$e_{ij} = \frac{o_{i.}\, o_{.j}}{n} = \frac{(i\text{-th row total})\,(j\text{-th column total})}{\text{number of observations}}$$

Time watching television	Sex		Row total
	Male: B_1	Female: B_2	
Over 25 hours: A_1	$\frac{(13)(15)}{30} = 6.5$	6.5	13
Under 25 hours: A_2	8.5	8.5	17
Column total	15	15	30

Now we calculate the χ^2 statistic as under

Category (i, j)	Observed frequency o_{ij}	Expected frequency e_{ij}	$\frac{(\lvert o_{ij} - e_{ij} \rvert - 0.5)^2}{e_{ij}}$
$A_1 B_1$	5	6.5	0.154
$A_1 B_2$	8	6.5	0.154
$A_2 B_1$	10	8.5	0.118
$A_2 B_2$	7	8.5	0.118
Total	30	30.0	$\chi^2 = 0.544$

Conclusion: Since $\chi^2 = 0.544 < 6.63$, we do not reject H_0: $\pi_{ij} = \pi_{i.}\,\pi_{.j}$ for all (i, j) against H_1: $\pi_{ij} \neq \pi_{i.}\,\pi_{.j}$ for at least one (i, j).

Exercise 15.2

1. (*a*) Define contingency table and cell frequency. What is a 2×2 contingency table.

(*b*) In an investigation into eye-colour and left or right handedness of a person, the following results were obtained:

	Handedness	
Eye colour	Left	Right
Blue	15	85
Brown	20	80

Do these results indicate, at the 5% level of significance, an association between eye colour and left or right handedness.

(Since $Adj\ \chi^2 = 0.56 < 3.84 = \chi^2_{1;\,0.95}$, we do not reject H_0: There is no association between eye colour and left or right handedness against H_1: There is association between eye colour and handedness.)

(*c*) An investigation into colour-blindness and sex of a person gave the following results:

	Colourblindness	
Sex	Colourblind	Not colourblind
Male	36	964
Female	19	981

Is there evidence, at the 5% level, of an association between the sex of a person and whether or not they are colourblind?

(Since $Adj\ \chi^2 = 4.79 > 3.84 = \chi^2_{1;\,0.95}$, we reject H_0: There is no association between sex of a person and colour-blindness in favour of H_1: There is association between sex of a person and colour-blindness.)

2. (*a*) A driving school examined the results of 100 candidates who were taking their driving test for the first time. They found that out of the 40 men, 28 passed and out of the 60 women, 34 passed. Do these results indicate, at the 5% level of significance, a

relationship between the sex of a candidate and the ability to pass first time ?

(Since $Adj\ \chi^2 = 1.290 < 3.84 = \chi^2_{1;0.95}$, we do not reject H_0: There is no relationship between the sex of a candidate and the ability to pass first time against H_1: There is relationship between the sex and ability to pass.)

(*b*) Out of 1350 persons, 450 were literate and 600 had traveled beyond the limits of their district, 300 of the literates were among those who had traveled. Find out by calculating (*i*) coefficient of association, (*ii*) the value of chi-square, if there is any association between traveling and literacy.

($Q = 0.6$, Since $Adj\ \chi^2 = 133.65 > 3.84 = \chi^2_{1;0.95}$, we reject H_0: There is no association between traveling and literacy in favour of H_1: There is association between traveling and literacy.)

3. (*a*) The following are the data on a random sample of 150 chickens, divided into two groups according to breed and into three group according to yield of eggs.

	Yield		
Breed	High	Medium	Low
Rhode Red	46	29	28
Leghorn White	27	14	6

Are these data consistent with the hypothesis that yield is not affected by the type of breed?

(Since $\chi^2 = 4.07 < 5.99 = \chi^2_{2;0.95}$, we do not reject H_0: There is no association between chicken breed and yield of eggs against H_1: There is association between chicken breed and yield of eggs.)

(*b*) The students of a college took three courses : arts, commerce and science. The students were classified according to the sex. The data on these students are given as follows :

	Course of study		
Sex	Arts	Commerce	Science
Male	200	300	100
Female	100	200	100

Use chi-square test whether there is any association between sex and choice of course of study.

(Since $\chi^2 = 13.888 > 5.99 = \chi^2_{2;0.95}$, we reject H_0: There is no association between sex and course of study in favour of H_1: There is association between sex and course of study.)

4. (*a*) The following table shows liking of three colours: pink, white and blue in samples of males and females:

	Sex	
Colour	Male	Female
Pink	20	40
White	40	20
Blue	60	20

Test whether there is any relation between sex and colour.

(Since $\chi^2 = 26.3889 > 5.99 = \chi^2_{2;\,0.95}$, we reject H_0: There is no relation between sex and liking of colours in favour of H_1: There is relation between sex and liking of colours.)

(*b*) The following table gives the condition at home and condition of the children.

Condition of children	Condition at home	
	Clean	Not clean
Clean	175	143
Fairly clean	136	116
Dirty	125	145

Test for the association between the conditions at home and condition of children.

(Since $\chi^2 = 5.027 < 5.99 = \chi^2_{2;\,0.95}$, we do not reject H_0: There is no association between conditions at home and condition of children against H_1: There is no association between conditions at home and condition of children.)

5. (*a*) The table given below shows the relation between the performance of students in economics and statistics. Test the hypothesis that the performance in economics is independent of the performance in statistics using 5% level of significance :

Grade in economics	Grade in statistics		
	High	Medium	Low
High	56	96	28
Medium	48	168	24
Low	16	86	78

(Since $\chi^2 = 89.2112 > 9.49 = \chi^2_{4;\,0.95}$, we reject H_0: There is no association between the performance of students in economics and statistics against H_1: There is association between the performance in economics and statistics.)

(*b*) A thousand households are taken at random and divided into three groups *A*, *B* and *C*, according to the total weekly income. The following table shows the numbers in each group having a colour television receive, a black and white receiver, or no television at all.

Television type	Income group		
	A	*B*	*C*
Colour television	56	51	93
Black and white	118	207	375
None	26	42	32

Calculate the expected frequencies if there is no association between total income and television ownership. Apply a test to find whether the observed frequencies suggest that there is such an association.

(Since $\chi^2 = 26.6 > 9.49 = \chi^2_{4;\,0.95}$, we reject H_0: There is no association between television type and income group in favour of H_1: There is association between television type and income group.)

6. (*a*) A random sample of 200 married men, all retired, were classified according to education and number of children as indicated below :

Education	Number of children		
	0 — 1	2 — 3	Over 3
Elementary	13	37	35
Secondary	19	42	14
College	12	17	11

Test the hypothesis, at 5% of significance, that the size of family is independent of the level of education attained by the father.

(Since $\chi^2 = 11.7194 > 9.49 = \chi^2_{4;\,0.95}$, we reject H_0: There is no association between education and number of children in favour, of H_1: There is association between education and number of children.)

(*b*) A survey of 200 families known to be regular television viewers was undertaken. They were asked which of the three television channel they watched most during an average week. A summary of their replies is given in the following table, together with the region in which they lived.

Channel	Region			
	North	East	South	West
PTV 1	29	16	42	23
PTV 2	6	11	26	7
STN	15	3	12	10

Test the hypothesis that there is no association between the channel watched most and the region.

(Since $\chi^2 = 13.446 > 12.59 = \chi^2_{6;\,0.95}$, we reject H_0: There is no association between the channel and region in favour of H_1: There is association between the channel and region.)

(*c*) From the following table showing the number of employees and condition of factory.

Condition of premises	Number of persons employed			
	Under 50	51 — 150	151 — 250	Over 250
A_1	84	133	49	62
A_2	87	82	20	25
A_3	26	9	9	5

Discuss the association between the condition of premises and the number of persons employed. Compute the coefficient of contingency.

(Since $\chi^2 = 30.06 > 12.59 = \chi^2_{6;\,0.95}$, we reject H_0: There is no association between the condition of premises and the number of persons employed in favour of H_1: There is association between the condition of premises and the number of persons employed. $C = 0.22$)

◆

15.7 RANK CORRELATION.

The correlation between ranks of individuals for both the variables X and Y is called *rank correlation*. A special case of correlation is when both the variables X and Y consist of sets of ranks. Suppose, for example, that two judges have ranked the same set of n objects according to some characteristic of interest. We are interested in determining whether the ranks assigned to the objects by one judge are related to or show any agreement with ranks assigned to the same objects by another judge.

15.7.1 Derivation of Spearman's Coefficient of Rank Correlation. Suppose that we have a sample of n individuals from a continuous bivariate population and two measurements for variables X and Y are made on each individual. We have n pairs of observations (a_1, b_1), $(a_2, b_2), \cdots, (a_n, b_n)$. These values for two variables can be ranked in separate ordered series. Let $x_1, x_2, \cdots, x_n$ be the ranks of $a_1, a_2, \cdots, a_n$ and $y_1, y_2, \cdots, y_n$ be the ranks of $b_1, b_2, \cdots, b_n$. The coefficient of rank correlation r_r is the ordinary correlation coefficient between the two sets of ranks. Then the coefficient of rank correlation is

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

The r_r always lies between -1 and $+1$. This formula is called Spearman's coefficient of rank correlation, in the honour of Charles Edward Spearman. Spearman's rank correlation coefficient is equivalent to Pearson's product moment correlation coefficient computed for ranks rather than the original observations. This nonparametric procedure can be useful in correlation analysis even when the basic data are not available in the form of numerical magnitudes but when the ranks can be assigned. The ranks may be assigned in order from high to low, with 1 representing the highest, 2 the next highest, *etc.* (or in order from low to high, with 1 representing the lowest, 2 the next lowest, *etc.*).

***Example* 15.9** *Using Spearman's formula calculate coefficient of rank correlation for the following data giving ranks to the measured quantities.*

a_i	4.7	2.9	6.4	2.5	4.9	7.3
b_i	8.6	5.4	6.2	4.9	8.3	7.2

Solution. The coefficient of rank correlation is obtained as

Measurements		Ranks			
a_i	b_i	x_i	y_i	$d_i = x_i - y_i$	d_i^2
4.7	8.6	4	1	3	9
2.9	5.4	5	5	0	0
6.4	6.2	2	4	−2	4
2.5	4.9	6	6	0	0
4.9	8.3	3	2	1	1
7.3	7.2	1	3	−2	4
Sum					18

$$r_r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(18)}{6\{(6)^2-1\}} = 0.4857$$

Exercise 15.3

1. (*a*) What is rank correlation?

(*b*) The following table shows how 10 students, arranged in alphabetical order, were ranked according to their achievements in both laboratory and lecture portions of a statistics course. Find the coefficient of mark correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

($r_r = 0.8545$)

(*c*) The ranks of the same 10 students in Mathematics and Economics were as follows:

(1, 6); (2, 5); (3, 1); (4, 4); (5, 2); (6, 7); (7, 8); (8, 10); (9, 3); (10, 9); the two numbers within brackets denoting the ranks of the same students in Mathematics, and Economics respectively. Calculate the rank correlation coefficient for proficiencies of this group in two subjects.

($r_r = 0.45$)

3. (*a*) Five sacks of coal *A*, *B*, *C*, *D* and *E* have different weights, with *A* being heavier than *B*, *B* being heavier than *C*, and so on. A weight lifter ranks the sacks (heaviest first) in the order *A*, *D*, *B*, *E*, *C*. Calculate a coefficient of rank correlation.

($r_r = 0.5$)

(*b*) Seven army recruits *A*, *B*, *C*, *D*, *E*, *F* and *G* were given two separate aptitude tests. Their orders of merit in each test were

Order of merit	1st	2nd	3rd	4th	5th	6th	7th
First test	*G*	*F*	*A*	*D*	*B*	*C*	*E*
Second test	*D*	*F*	*E*	*B*	*G*	*C*	*A*

Find Spearman's coefficient of rank correlation between the two orders and comment briefly on the correlation obtained.

($r_r = -0.036$, Very little negative correlation)

4. (*a*) Ten competitors in a beauty contest are ranked by three judges in the following order

Competitor	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Judge *X*	1	6	5	10	3	2	4	9	7	8
Judge *Y*	3	5	8	4	7	10	2	1	6	9
Judge *Z*	6	4	9	8	1	2	3	10	5	7

Use Spearman's rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

($r_{xy} = -0.21$, $r_{yz} = -0.30$, $r_{xz} = 0.64$; This indicates that judges the *X* and *Z* have the nearest approach to common tastes in beauty.)

(*b*) The following table shows the grade point average awarded to six children in a competition by two different judges.

Child	*A*	*B*	*C*	*D*	*E*	*F*
Judge *X*	6.8	7.3	8.1	9.8	7.1	9.2
Judge *Y*	7.8	9.4	7.9	9.6	8.9	6.9

Calculate coefficient of rank correlation by Spearman's formula.
(r_r = 0.26)

(*c*) The following table shows the marks of six candidates in two subjects.

Candidate		*A*	*B*	*C*	*D*	*E*	*F*
Mathematics	x_i	38	62	56	42	59	48
Statistics	y_i	64	89	84	60	73	69

(*i*) Calculate the coefficient of rank correlation.

(*ii*) Comment on the value of your result.

{ (*i*) 0.886, (*ii*) High positive correlation }

Exercise 15.4
Objective Questions

1. Fill in the blanks.

(*i*) A characteristic which varies in quantity from one individual to another is called a ———. (*variable*)

(*ii*) A characteristic which varies in quality from one individual to another is called an ———. (*attribute*)

(*iii*) The observations made on objects regarding an attribute are called ——— data. (*qualitative*)

(*iv*) ——— is a process of dividing the objects into two mutually exclusive classes of an attribute. (*Dichotomy*)

(*v*) The degree of linear relationship between the two variables is called ———. (*correlation*)

(*vi*) The degree of relationship between the two attributes is called ———. (*association*)

(*vii*) The two attributes A and B are ———, if

$$(AB) = \frac{(A)(B)}{n}$$ (*independent*)

(*viii*) The two attributes A and B are ———, if

$$(AB) \neq \frac{(A)(B)}{n}$$ (*associated*)

(*ix*) The two attributes A and B are ———— associated, if

$$(AB) > \frac{(A)(B)}{n}$$ (*positively*)

(*x*) The two attributes A and B are ———— associated, if

$$(AB) < \frac{(A)(B)}{n}$$ (*negatively*)

(*xi*) The coefficient of association, denoted by Q, is a measure of association between the two ————. (*attributes*)

(*xii*) If the coefficient of association equals 0, the two attributes A and B are ————. (*independent*)

(*xiii*) If the coefficient of association is not equal to 0, the two attributes A and B are ————. (*associated*)

(*xiv*) If the coefficient of association equals -1, the two attributes A and B are completely ————. (*dissociated*)

(*xv*) If the coefficient of association equals 1, the two attributes A and B are completely ————. (*associated*)

(*xvi*) A ———— table consisting of r rows and c columns is made up of the observed frequencies relative to two attributes and their categories. (*contingency*)

(*xvii*) The two attributes are said to be ————, if for every cell of a contingency table the observed frequency o_{ij} is equal to expected frequency e_{ij}. (*independent*)

(*xx*) For an $r \times c$ contingency table, the χ^2-statistic has degrees of freedom ν = ————. $(r-1)(c-1)$

(*xxiv*) The larger are the difference between the observed and expected frequencies, the larger will be the value of χ^2 which leads to the ———— of H_0 of independence. (*rejection*)

(*xxv*) The rejection of H_0 of independence indicates that the two criteria of classification are ————. (*associated*)

(*xxvi*) In a chi-square test for independence, no expected frequency should be ———— than 5. (*less*)

2. Mark off the following statements as false or true.

(*i*) A characteristic which varies in quantity from one individual to another is called an attribute. (*false*)

(*ii*) The quantitative data relating to an attribute may be obtained simply by noting its presence or absence in the objects. (*true*)

(*iii*) The presence of attributes is denoted by capital Latin letters and their absence by Greek or small letters. (*true*)

(*iv*) The class frequencies of the highest order are called ultimate class frequencies. (*true*)

(*v*) The two attributes A and B are associated, if.

$$(AB) = \frac{(A)(B)}{n}$$ (*false*)

(*vi*) The two attributes A and B are positively associated, if

$$(AB) < \frac{(A)(B)}{n}$$ (*false*)

(*vii*) The coefficient of correlation, denoted by r, is a measure of the strength of linear relationship between two variables. (*true*)

(*viii*) The coefficient of association, denoted by Q, is a measure of association between the two attributes. (*true*)

(*ix*) The coefficient of association always lies between -1 and 1. (*true*)

(*x*) A contingency table consisting of r rows and c columns is made up of the observed frequencies relative to two attributes and their categories. (*true*)

(*xi*) The disassociation of two attributes means their independence (*false*)

(*xii*) A measure of the discrepancy between the observed and expected frequencies is called a chi-square (χ^2) test of independence. (*true*)

(*xiii*) The value of χ^2-statistic is always non-negative. (*true*)

(*xiv*) The larger are the differences between the observed and expected frequencies, the larger will be the value of χ^2 which leads to rejection of H_0 of independence. (*true*)

(*xv*) The rejection of H_0 of independence indicates that two criteria of classification are associated. (*true*)

16 ANALYSIS OF TIME SERIES

16.1 TIME SERIES

The sequence $y_1, y_2, \cdots, y_n$ of n observations of a variable Y, recorded in accordance with their time of occurrence $t_1, t_2, \cdots, t_n$, is called a *time series.* Symbolically, the variable Y can be expressed as a function of time t as

$$y = f(t) + e$$

where $f(t)$ is a completely determined or specified sequence that follows a systematic pattern of variation and e is a random error that follows an irregular pattern of variation.

Signal. The signal is a systematic component of variation in a time series.

Noise. The noise is an irregular component of variation in a time series.

Therefore, a time series is a sequence of observations, on a variable, that are arranged in chronological order. The observations in a time series are usually made at equidistant points of time. Examples of a time series are: the hourly temperature recorded at a weather bureau, the total annual yield of wheat over a number of years, the monthly sales of a fertilizer at a store, the enrolment of students in various years in a college, the daily sales at a departmental store, *etc.*

16.1.1 Historigram. A *historigram* is a graphic representation of a time series that reveals the changes occurred at different time periods. A first step in the prediction or forecast of a time series involves an examination of the set of past observations. The construction of a historigram involves the following steps:

(*i*) Using an appropriate scale, take time t along *x-axis* as an independent variable.

(*ii*) Using an appropriate scale, plot the observed values of variable Y as a dependent variable against the given points of time.

(*iii*) Join the plotted points by line segments to get the required historigram.

***Example* 16.1** *Draw a historigram to show the population of Pakistan in various census years*

Census year	Population (million)
1951	33.44
1961	42.88
1972	65.31
1981	83.78
1998	130.58

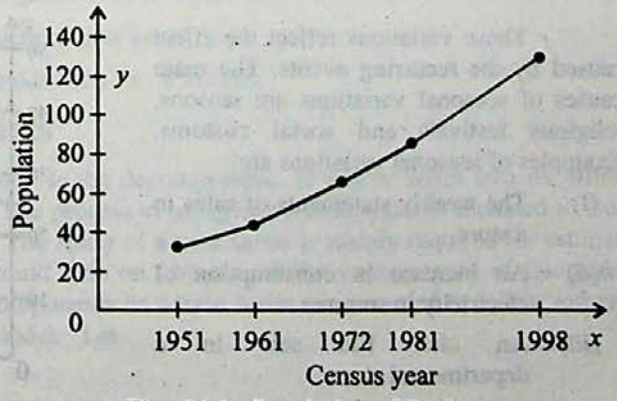


Fig. 16.1 Population of Pakistan

Solution. The population of Pakistan in different census years is represented by a historigram as shown in Fig. 16.1.

16.2 COMPONENTS OF A TIME SERIES

The examples of time series suggest that a typical time series may be composed of the following four components:

(*i*) Secular trend (T)

(*ii*) Seasonal variations (S)

(*iii*) Cyclical fluctuations (C)

(*iv*) Irregular movements (I)

These are the basic components of a time series, each of which is regarded as the result of a well defined distinct cause. A time series is not necessarily composed of all these four components.

16.2.1 Secular (Long-term) Trend. The *secular trend* is a line or curve that shows the general tendency of a time series. It represents a relatively smooth, steady, and gradual movement of a time series in the same direction (upward or downward). It shows the general increase or decrease in a sequence of observations, and reflects the effect of the forces operating over a fairly long period of time. Examples of secular trend are:

(*i*) The decline in death rate due to advancement in science.

(*ii*) A continually increasing demand for smaller automobiles.

(*iii*) A need for increased wheat production due to a constant increase in population.

16.2.2 Seasonal Variations. The *seasonal variations* are short term movements that represent the regularly recurring changes in a time series. These variations indicate a repeated pattern of identical changes in the data that tend to recur regularly during a period of one year or less. These changes are repeated with the same pattern within a specific time period, called the *periodicity*. Seasonal variations may have the fixed periodicity, such as daily, weekly, monthly, or yearly *etc.* These changes are periodic in nature and their influence; upon a specific time series is fairly regular, both in respect of length (*time*) and amplitude (*size*).

These variations reflect the effect caused by the recurring events. The main causes of seasonal variations are seasons, religious festivals and social customs. Examples of seasonal variations are:

(*i*) The weekly statements of sales in a store.

(*ii*) An increase in consumption of electricity in summer.

(*iii*) An after Eid sale in a departmental store.

(*iv*) An increase in sales of cold drinks during summer.

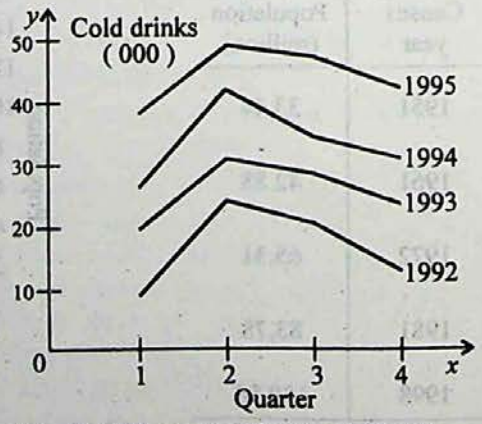


Fig. 16.2 The seasonal pattern of cold drinks sales

16.2.3 Cyclical Fluctuations. The *cyclical fluctuations* are the long term oscillations about the trend. These are the periodic up-and-down movements in a time series that tend to recur over a long period of time. The cyclic patterns tend to vary in length (time) and amplitude (size) and they are differentiated from the seasonal variations by the fact that they do not have a fixed periodicity. Although, these variations are recurring yet are less predictable than seasonal variations and secular trend, therefore, they have a more dangerous effect on a business and economic activity. These fluctuations reflect the effect caused by a so called *business cycle*. A business cycle has the following four phases:

(*i*) Trough (Depression)

(*ii*) Expansion (Recovery)

(*iii*) Peak (Boom or Prosperity)

(*iv*) Recession (Contraction)

A *trough* is the lowest point relative to the rest of the particular cycle. After the downswing has run its course, the *expansion* phase reverses direction and starts rising. The upswing eventually levels off and reaches its *peak*. This is the highest point relative to the particular cycle. Finally, the upswing starts to turn downward. We refer to this following phase as a *recession*.

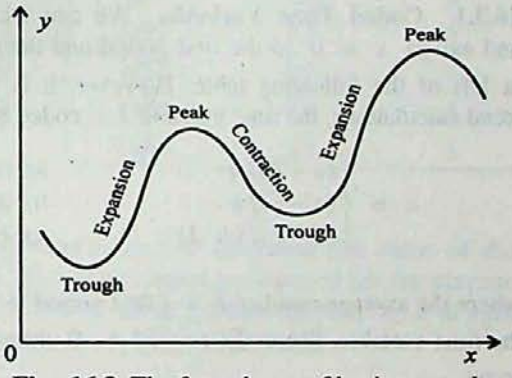


Fig. 16.3 The four phases of business cycle

16.2.4 Irregular Movements. The *irregular movements* are unpredictable changes that indicate the effect of random events. The examples of random events are wars, floods, earthquakes, strikes, fires, elections *etc*. The irregular movements are unsystematic, non-recurring, accidental and unusual in nature. These variations are also known as *erratic, accidental* or *random* variations. Examples of irregular movements are:

(*i*) A steel strike, delaying production for a week.

(*ii*) A fire in a factory delaying production for 3 weeks.

16.3 ANALYSIS OF TIME SERIES

The *analysis of a time series* is the decomposition of a time series into its different components for their separate study. The process of analysing a time series is intended to isolate and measure its various components. The study of a time series is mainly required for estimation and forecasting. An ideal forecast should base on forecasts of the various types of fluctuations. While performing the analysis, the components of a time series are assumed to follow either the multiplicative model or the additive model. Let

Y = Original observation,

T = Trend component, S = Seasonal component,

C = Cyclical component, I = Irregular component.

In the *multiplicative model*, it is assumed that the value Y of a composite series is the product of the four components T, S, C and I. Symbolically,

$$Y = T \times S \times C \times I$$

where the component T is given in original units of Y but the other components S, C and I are expressed as percentage unitless index numbers.

In the *additive model*, it is assumed that the value Y of the composite series is the sum of the four components T, S, C and I. Symbolically,

$$Y = T + S + C + I$$

where the components T, S, C and I all are given in the original units of Y. Conventionally, the multiplicative model is considered as the standard model for analysis of a time series.

16.3.1 Coded Time Variable. We can take the origin at the beginning of a time series and assign $x = 0$ to the first period and then number other periods as $1, 2, 3, \cdots$ as shown at left of the following table. However, it is important to note that in order to simplify the trend calculations, the time variable t is coded by

$$x = \begin{cases} (t - \bar{t})/h & \text{for odd number of periods} \\ (t - \bar{t})/h & \text{for even number of periods in units of } h \text{ period} \\ (t - \bar{t})/(h/2) & \text{for even number of periods in units of } h/2 \text{ period} \end{cases}$$

where the average number $\bar{t} = (\text{first period} + \text{last period})/2$ and h is the constant interval in the time variable. Since $\sum(t - \bar{t}) = 0$, then we get $\sum x = 0 = \sum x^3 = \sum x^5 = \cdots$, and so on.

The odd number of years in period 1980 — 1984 at the middle of the following table has $\bar{t} = (1980 + 1984)/2 = 1982$ as the middle point. The code for the year t is $x = t - \bar{t}$. For $t = 1982$, we have $x = t - \bar{t} = 1982 - 1982 = 0$. Thus, the coded year is zero at $\bar{t}$. For $t = 1980$, we have $x = 1980 - 1982 = -2$. Actually, the only computation we need is that for $\bar{t}$. Thus after entering $x = 0$ at the middle of an odd number of years, we assign $-1, -2, \cdots$ and so on for the years before the middle year, and $1, 2, \cdots$ and so on for the years after the middle year as shown in the following table.

The even number of years in period 1980 — 1985 at the right of the following table has $\bar{t} = (1980 + 1985)/2 = 1982.5$ as the middle point. So $x = 0$ half way between the years 1982 and 1983. For $t = 1982$, we have $x = t - \bar{t} = 1982 - 1982.5 = -0.5$. Then after considering $x = 0$ at the middle of an even number of years, we assign $-0.5, -1.5, -2.5, \cdots$ and so on for the years before the middle year and $0.5, 1.5, 2.5, \cdots$ and so on for the years, after the middle year as shown in the following table.

However, to avoid decimals in the coded years we can take the unit of measurement as 1/2 year. Then after considering $x = 0$ at the middle of an even number of years, we assign $-1, -3, -5, \cdots$ and so on for the years before the middle year and $1, 3, 5, \cdots$ and so on for the years after the middle year as shown in the following table.

Table: Coded Year Number

Origin at beginning. The starting year is coded $x = 0$		Odd number of years. The middle year is coded $x = 0$		Even number of years. $x = 0$ is at the centre of two middle years		
Year	Coded year in units as one year	Year	Coded year in units as one year	Year	Coded year in units as one year	Coded year in units as 1/2 year
t	x	t	$x = t - \bar{t}$	t	$x = t - \bar{t}$	$x = \frac{t - \bar{t}}{1/2}$
1980	0	1980	−2	1980	−2.5	−5
1981	1	1981	−1	1981	−1.5	−3
1982	2	1982	0	1982	−0.5	−1
1983	3	1983	1	1983	0.5	1
1984	4	1984	2	1984	1.5	3
				1985	2.5	5

16.4 ESTIMATION OF SECULAR TREND

It has been earlier stated that one component force that determine the value of the variable at any period of time is the secular trend. The secular trend is measured for the purpose of prediction or projection into the future. The secular trend can be represented either by a straight line or by some type of smooth curve. It is measured by the following methods:

(*i*) Method of free hand curve

(*ii*) Method of semi-averages

(*iii*) Method of moving averages

(*iv*) Method of least squares

16.4.1 Method of free hand curve. The secular trend is measured by the method of free hand curve in the following steps.

(*i*) Using an appropriate scale, take the time periods along *x-axis*, as an independent variable.

(*ii*) Using an appropriate scale, plot the points for observed values of the variable Y as a dependent variable against the given time periods.

(*iii*) Join these plotted points by line segments to get a historigram.

(*iv*) Keeping in view the up and down fluctuations of the graph, draw a free hand smooth curve or a straight line through the historigram in a way such that it indicates the general trend of the time series.

(*v*) Instead of locating the line simply by eye looking at the graph, the average $\bar{y}$ of original values may be used as the trend value $\bar{y}'$ at the middle of the time period. Plot this average in the middle of the time period and the required trend line or curve should be drawn through this point, as it is a reasonable condition that $\bar{y}$ should be equal to $\bar{y}'$.

(*vi*) Read off the trend values for different time periods from this trend line or curve.

If a straight line is used for locating the trend, then it becomes easy to estimate the rate of change (slope of the line b) by measuring the difference $y'_{x+1} - y'_x$ of the trend values for any two consecutive time periods x and $x + 1$. Symbolically it is expressed as $b = y'_{x+1} - y'_x$. Then the equation of the trend line is summarised in the slope intercept from as $y' = a + bx$ with origin at any time period, so that, a = trend value for the origin.

If the historigram indicates a non-linear trend, then in such situations it is generally preferred to use a curve instead of a straight line to show the secular trend.

Merits.

(*i*) The free hand curve method is a simple, easy and quick method for measuring secular trend.

(*ii*) The trend line or curve smoothes out seasonal variations.

(*iii*) A good fitted trend line or curve can give a close approximation to the trend based on a mathematical model.

Demerits.

(*i*) It is a rough and crude method. It is greatly affected by the personal bias, *i. e.*, different persons may fit different trends to the same data.

(*ii*) It requires too much practice to get a good fit.

(*iii*) The free hand curve method is subject to personal bias, so it is unable to give reliable estimates.

***Example* 16.2** *The following time series shows the number of road accidents in Punjab for the year* 1972 *to* 1978.

Year	1972	1973	1974	1975	1976	1977	1978
Number of accidents	2493	2638	2699	3038	3745	4079	4688

(*i*) *Obtain the historigram showing the number of road accidents and a free hand trend line by drawing a straight line.*

(*ii*) *Find the trend values for this time series.*

Solution. (*i*)

Year	Value y	Total	Mean	Trend value
1972	2493			2200
1973	2638			2550
1974	2699			2950
1975	3038	23380	3340	3340
1976	3745			3650
1977	4079			4050
1978	4688			4400

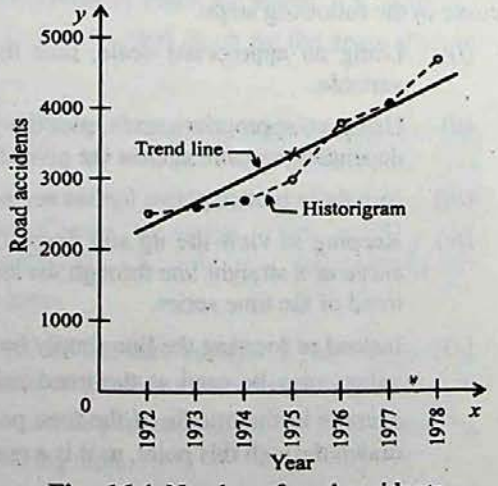


Fig. 16.4 Number of road accidents

(*ii*) Reading off the trend line, we get the trend values.

16.4.2 Method of Semi-averages. The secular trend is measured by the method of semi-averages in the following steps.

(*i*) Divide the observed values of the time series into two equal periods. If the number of observed values is odd then it is advisable either to omit the middle value altogether or to include the middle value in each half.

(*ii*) Take the average of each part and place these average values against the mid points of the two parts, and the average value of each part should be considered equal to the average value of its respective trend values.

(*iii*) Plot the semi-averages on the graph of the original values.

(*iv*) Draw the required trend line through these two plotted points, and extend it to cover the whole period.

(*v*) With two points located on the straight line, it is simple to compute the slope and *y*-intercept of the line. This slope gives the estimate of the rate of change of values. Now, the trend values are found either by reading off the semi-average trend line or by the estimated straight line as explained below

Semi-average Trend Line. Let y_1' and y_2' be the semi averages placed against the times x_1 and x_2, and the estimated straight line (in slope-intercept from) $y' = a + b\,x$ is to pass through the points (x_1, y_1') and (x_2, y_2'). The constants a (y-intercept) and b (slope of the line) can be easily determined. The equation of the line passing through the points (x_1, y_1') and (x_2, y_2') can be written as

$$y' - y_1' = \frac{y_2' - y_1'}{x_2 - x_1}(x - x_1)$$

$$y' - y_1' = b(x - x_1) \qquad \text{where} \quad b = \frac{y_2' - y_1'}{x_2 - x_1}$$

$$y' = (y_1' - b\,x_1) + b\,x$$

$$y' = a + b\,x \qquad \text{where} \quad a = y_1' - b\,x_1$$

If the number of time units in the observed time series is even, then the following formula may be used to find the slope of the trend line.

$$b = \frac{1}{n/2}\left(\frac{S_2}{n/2} - \frac{S_1}{n/2}\right)$$

$$= \frac{1}{n/2}\left(\frac{S_2 - S_1}{n/2}\right) = \frac{(S_2 - S_1)}{n^2/4}$$

$$= \frac{4(S_2 - S_1)}{n^2}$$

where S_1 = sum of y-values for the first half of the period.

S_2 = sum of y-values for the second half of the period.

n = number of time units covered by the time series.

Merits.

(*i*) The method of semi-averages is simple, easy and quick.

(*ii*) It gives an objective result.

(*iii*) It smoothes out seasonal variations.

(*iv*) It gives better approximation to the trend because it is based on a mathematical model as compared to free hand method.

Demerits.

(*i*) The arithmetic mean, which is used to average the two halves of the observed values, is highly affected by extreme values.

(*ii*) This method can only be applied if the trend is linear or approximately linear.

(*iii*) This method is not appropriate if the trend is not linear.

***Example* 16.3** *The following table shows the property damaged by road accidents in Punjab for the years* 1973 *to* 1979.

Year	1973	1974	1975	1976	1977	1978	1979
Property damaged	201	238	392	507	484	649	742

(*i*) *Obtain the semi-averages trend line.*

(*ii*) *Find out the trend values.*

Solution. (*i*) Let $x = t - 1973$.

Year t	Property damaged y	Semi-total	Semi-average	Coded year $x = t - 1973$	Trend value $y' = 190 + 87x$
1973	201	831	277	0	$190 + 87(0) = 190$
1974	238			1	$190 + 87(1) = 277$
1975	392			2	$190 + 87(2) = 364$
1976	507			3	$190 + 87(3) = 451$
1977	484	1875	625	4	$190 + 87(4) = 538$
1978	649			5	$190 + 87(5) = 625$
1979	742			6	$190 + 87(6) = 712$

The semi averages trend line is

$$y' = a + bx$$

Taking the origin at 1973, we have

$y_1' = 277, \quad x_1 = 1$

$y_2' = 625, \quad x_2 = 5$

$$b = \frac{y_2' - y_1'}{x_2 - x_1}$$

$$= \frac{625 - 277}{5 - 1} = 87$$

$$a = y_1' - b x_1$$

$$= 277 - 87(1) = 190$$

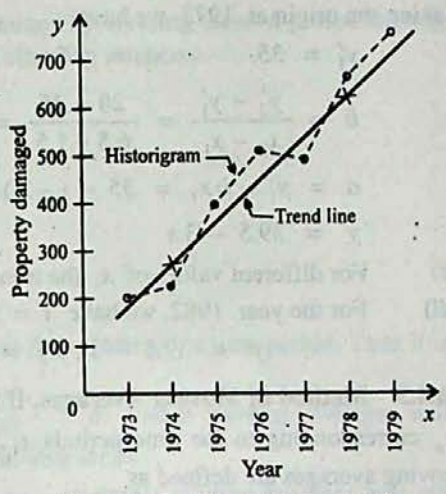


Fig. 16.5 Property damaged

The semi-averages trend line is

$y' = 190 + 87x$ with origin at 1973

(ii) For different values of x, the trend values are obtained as shown in the table.

***Example* 16.4** *The following table gives the number of books (in 000's) sold at a book stall for the year* 1973 *to* 1981.

Year	1973	1974	1975	1976	1977	1978	1979	1980	1981
Number of books (000's)	42	38	35	25	32	24	20	19	17

(*i*) *Find the equation of the semi-averages trend line.*

(*ii*) *Compute out the trend values.*

(*iii*) *Estimate the number of books sold for the year* 1982.

Solution. (*i*) Let $x = t - 1973$.

Year t	No. of books y	Semi-total	Semi-average	Coded year $x = t - 1973$	Trend value $y' = 39.5 - 3x$
1973	42	140	35	0	$39.5 - 3(0) = 39.5$
1974	38			1	$39.5 - 3(1) = 36.5$
1975	35			2	$39.5 - 3(2) = 33.5$
1976	25			3	$39.5 - 3(3) = 30.5$
1977	32			4	$39.5 - 3(4) = 27.5$
1978	24	80	20	5	$39.5 - 3(5) = 24.5$
1979	20			6	$39.5 - 3(6) = 21.5$
1980	19			7	$39.5 - 3(7) = 18.5$
1981	17			8	$39.5 - 3(8) = 15.5$

The semi-averages trend line is

$y' = a + bx$

Taking the origin at 1973, we have

$$y_1' = 35, \qquad x_1 = 1.5 \qquad \text{and} \qquad y_2' = 20, \qquad x_2 = 6.5,$$

$$b = \frac{y_2' - y_1'}{x_2 - x_1} = \frac{20 - 35}{6.5 - 1.5} = -3$$

$$a = y_1' - b x_1 = 35 - (-3)(1.5) = 39.5$$

$$y' = 39.5 - 3x \qquad \text{with origin at 1973}$$

(*ii*) For different values of x, the trend values are obtained as shown in the table.

(*iii*) For the year 1982, we have $x = 1982 - 1973 = 9$. Then

$$y' = 39.5 - 3(9) = 12.5$$

16.4.3 Method of Moving Averages. If the observed values of a variable Y are $y_1, y_2, \cdots, y_n$ corresponding to the time periods $t_1, t_2, \cdots, t_n$ respectively, then the k-period simple moving averages are defined as

$$a_1 = \frac{1}{k}\sum_{i=1}^{k} y_i, \qquad a_2 = \frac{1}{k}\sum_{i=2}^{k+1} y_i,$$

$$a_3 = \frac{1}{k}\sum_{i=3}^{k+2} y_i, \qquad \cdots, \qquad a_m = \frac{1}{k}\sum_{i=m}^{n} y_i,$$

where $a_1, a_2, a_3, \cdots, a_m$ is the sequence of k-period simple moving averages. That is, the k-period simple moving averages are calculated by averaging first k observations and then repeating this process of averaging the k observations by dropping each time the first observation and including the nest one that has not been previously included. This process is continued till the last k observations have been averaged. For example, the 3-period simple moving averages are given as

$$a_1 = \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}\sum_{i=1}^{3} y_i$$

$$a_2 = \frac{1}{3}(y_2 + y_3 + y_4) = \frac{1}{3}\sum_{i=2}^{4} y_i$$

$$a_3 = \frac{1}{3}(y_3 + y_4 + y_5) = \frac{1}{3}\sum_{i=3}^{5} y_i$$

and so on. Each of these simple moving average of the sequence $a_1, a_2, a_3, \cdots$ is placed against the middle of each successive group. For practical purposes the k-period moving successive totals $S_1, S_2, S_3 \cdots$ are obtained by the following relations

$$S_1 = \sum_{i=1}^{k} y_i$$

$$S_2 = S_1 + y_{k+1} - y_1$$

$$S_3 = S_2 + y_{k+2} - y_2$$

and so on. The k-period simple moving averages are obtained by dividing these k-period moving successive totals $S_1, S_2, S_3 \cdots$ by k as given in the following relations.

$$a_1 = \frac{S_1}{k}$$

$$a_2 = a_1 + \frac{y_{k+1} - y_1}{k}$$

$$a_3 = a_2 + \frac{y_{k+2} - y_2}{k}$$

and so on. Each moving average should be placed against the middle of its time period. Then it is obvious that

(*i*) When k is odd, the sequence $a_1, a_2, a_3, \cdots$ of simple moving averages will correspond directly to the observed values in the time series.

(*ii*) When k is even, the sequence $a_1, a_2, a_3, \cdots$ of simple moving averages will not correspond directly to the observed values in the time series and will be placed in the middle of two time periods. It is then sometimes necessary to centralize these averages so that they should correspond to the observed values in the time series. For centralization, further 2-period moving averages of the former k-period moving averages are computed which are called k-period centred moving averages.

Smoothing of a Time Series. The smoothing of a time series is a process of eliminating the unwanted fluctuations in a time series. The moving averages tend to reduce the variation present among the observed values of a time series, so they are used to eliminate the unwanted fluctuations. Thus the moving averages may be used in smoothing of a time series. They eliminate the effect of periodic fluctuations if an appropriate period moving averages are calculated. For this purpose the period of the moving average is chosen such that it should be equal to the period of at least one cycle. The secular trend is measured by taking the following steps.

(*i*) Find the moving averages of an appropriate period.

(*ii*) Plot the points representing these moving averages on the graph of the observed time series and join these points by the line segments.

(*iii*) The graph of the moving averages indicates the secular trend by eliminating the periodic fluctuations

The period of moving averages should be decided in the light of the periodicity of a time series. Because only the moving averages, calculated by using the time period which approximately coincides with the periodicity of the time series, would eliminate, nearly completely, all its regular fluctuations and show a trend.

Merits.

(*i*) The method of moving averages is easy and simple.

(*ii*) The moving averages of an appropriate period eliminate the periodic fluctuations, so it may be used to eliminate cyclical and seasonal fluctuations.

Demerits.

(*i*) The method of moving averages does not give the trend values at the beginning and at the end of the original time series.

(*ii*) The moving averages are highly affected by the extreme observations however the affect may be reduced by using the geometric mean as average.

(*iii*) The method of moving averages does not provide a mathematical equation for the trend, therefore, the forecasting is only subjective.

(*iv*) The selection of inappropriate period of moving averages may generate the cycles which are not present in the observed time series.

***Example* 16.5** *The following table shows the production of silver utensils (in thousands) at a certain factory in Gujranwala.*

Year	Utensils (000)
1970	170.0
1971	154.8
1972	156.5
1973	158.9
1974	140.3
1975	154.2
1976	160.7
1977	178.3

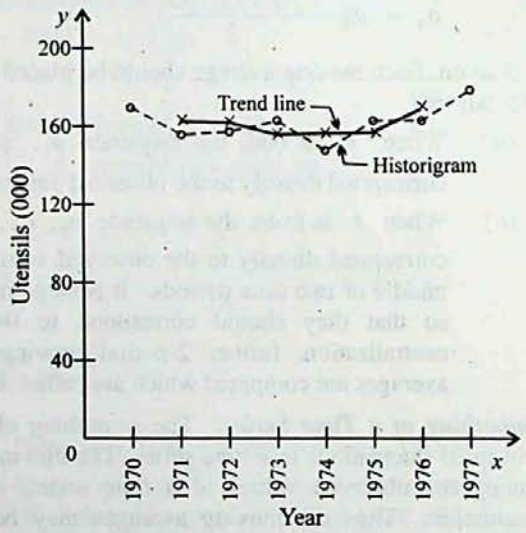


Fig. 16.6 Production of silver utensils

(*i*) *Calculate* 3 *year simple moving averages for the following time series.*

(*ii*) *Also plot actual data and moving averages on a graph.*

Solution.

Year	Production y	3-year moving total	3-year moving average
1970	170.0		
1971	154.8	481.3	160.43
1972	156.5	470.2	156.73
1973	158.9	455.7	151.90
1974	140.3	453.4	151.13
1975	154.2	455.2	151.73
1976	160.7	493.2	164.40
1977	178.3		

***Example* 16.6** *The following table shows the food grain price index number of quarters for the years* 1962 *and* 1963.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1962	93	97	96	93
1963	97	102	106	98

Calculate four quarter moving average centred.

Solution. The four quarter centred moving averages are obtained as under:

(1) Year	(2) Quarter	(3) Price index number y	(4) 4-quarter moving total	(5) 4-quarter moving average (4) ÷ 4	(6) 2-quarter moving total of (5)	(7) 4-quarter centred moving average (6) ÷ 2
1962	I	93				
	II	97	379	94.75		
	III	96			190.50	95.25
	IV	93	383	95.75		
1963	I	97			192.75	96.38
	II	102	388	97.00		
	III	106			196.50	98.25
	IV	98	398	99.50		
					200.25	100.12
			403	100.75		

Alternately, for the sake of convince, the four quarter centred moving averages may be calculated as shown in the table given below:

(1) Year	(2) Quarter	(3) Price index number y	(4) 4-quarter moving total	(5) 4-quarter centred moving total	(6) 4-quarter centred moving average (5) ÷ 8
1962	I	93			
	II	97	379		
	III	96		762	95.25
	IV	93	383	771	96.38
1963	I	97	388	786	98.25
	II	102	398	801	100.12
	III	106	403		
	IV	98			

16.4.4 Method of Least Squares. For situations in which it is desirable to have a mathematical equation to describe the secular trend of a time series, the most commonly used method is to fit a straight line $\hat{y} = a + bx$, a second degree parabola $\hat{y} = a + bx + cx^2$, *etc.*, where y is the value of a time series variable, x representing the time and all others are constants. For determining the values of the constants appearing in such an equation, the most widely used method is the method of least squares, because it is a practical method that provides best fit according to a reasonable criterion. The principle of least squares says that "*the sum of squares of the deviations of the observed values from the corresponding expected values should be least*".

Among all the trend lines approximating a given time series data, the trend line is called a least squares fit for which the sum of the squares of the deviations of the observed values from

their corresponding expected values is the least. The method of least squares consists of minimizing the sum of the squares of these deviations. To avoid the personal bias in measuring the secular trend this method is used to find a trend line approximating a given time series.

Secular Trend — Linear. It is useful to describe the trend in a time series where the amount of change is constant per unit time.

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be the n pairs of observed sample values of a time series variable y, with x representing the coded time value. We can plot these n points on a graph. Because of the fact that $y_1, y_2, \cdots, y_n$ are observed values of a time series variable, these points will not necessarily lie on a straight line. Let us suppose that we want to fit a straight line expressed in slope-intercept form as

$$\hat{y} = a + bx$$

This line will be called the least squares line if it makes $\sum (y - a - bx)^2$ minimum. The method of least squares yields the following normal equations.

$$\sum y = na + b\sum x, \qquad \sum xy = a\sum x + b\sum x^2$$

The normal equations give the values of a and b as

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}, \qquad a = \frac{\sum y - b\sum x}{n} = \bar{y} - b\bar{x}$$

However, if $\sum x = 0$, then the usual normal equations reduce to

$$\sum y = na, \qquad \sum xy = b\sum x^2$$

Therefore, the values of a and b also reduce to

$$a = \frac{\sum y}{n} = \bar{y}, \qquad b = \frac{\sum xy}{\sum x^2}$$

The trend values $\hat{y}$ are computed from the least squares line $\hat{y} = a + bx$ by substituting the values of x corresponding to the different time periods. The secular trend can be indicated on a graph by plotting these estimated vales against their respective time periods.

***Properties*:**

(*i*) The least squares line always passes through the point $(\bar{x}, \bar{y})$ called the centre of gravity of the data.

(*ii*) The sum of the deviations $\sum(y - \hat{y})$ of the observed values y from their corresponding expected values $\hat{y}$ is zero, *i. e.*,

$$\sum(y - \hat{y}) = 0 \quad \Rightarrow \quad \sum y = \sum \hat{y}$$

(*iii*) The sum of squares of the deviations $\sum(y - \hat{y})^2$ measures how well the trend line fits the data. A smaller $\sum(y - \hat{y})^2$ means the better fit.

***Example* 16.7** *The following table shows the production of steel in a steel mill for the time period* 1977 *to* 1983.

Year	1977	1978	1979	1980	1981	1982	1983
Production (000 tons)	12.7	10.1	13.0	13.2	12.6	14.2	13.7

Find the linear trend by the method of least squares by taking the origin:

(*i*) *at the beginning period of the time period,*

(*ii*) *at the middle of the time period* 1977 — 83.

Calculate the trend values in both cases.

Solution. (*i*) Taking the origin at the beginning period, 1977 (*i. e.*, July 1, 1977), we have $x = t - 1977$.

Year t	Production y	Coded year $x = t - 1977$	xy	x^2	Trend value $\hat{y} = 11.628 + 0.386\,x$
1977	12.7	0	0	0	11.628 + 0.386 (0) = 11.628
1978	10.1	1	10.1	1	11.628 + 0.386 (1) = 12.014
1979	13.0	2	26.0	4	11.628 + 0.386 (2) = 12.400
1980	13.2	3	39.6	9	11.628 + 0.386 (3) = 12.786
1981	12.6	4	50.4	16	11.628 + 0.386 (4) = 13.172
1982	14.2	5	71.0	25	11.628 + 0.386 (5) = 13.558
1983	13.7	6	82.2	36	11.628 + 0.386 (6) = 13.944
Total	89.5	21	279.3	91	

The least squares trend line is

$$\hat{y} = a + b\,x$$

The least squares estimates a and b are

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{7(279.3) - (21)(89.5)}{7(91) - (21)^2} = 0.386$$

$$a = \frac{\sum y - b\sum x}{n} = \frac{89.5 - (0.386)(21)}{7} = 11.628$$

The best fitted line is

$$\hat{y} = 11.628 + 0.386\,x \qquad \text{with origin at 1977}$$

For different values of x, the trend values are obtained as shown in the table.

(*ii*) We have $\bar{t} = (1977 + 1983)/2 = 1980$. Taking the origin at the middle of the time period at 1980 (*i. e.*, July 1, 1980), we have $x = t - \bar{t} = t - 1980$.

Year t	Production y	Coded year $x = t - 1980$	xy	x^2	Trend values $\hat{y} = 12.786 + 0.386\,x$
1977	12.7	−3	−38.1	9	12.786 + 0.386 (−3) = 11.628
1978	10.1	−2	−20.2	4	12.786 + 0.386 (−2) = 12.014
1979	13.0	−1	−13.0	1	12.786 + 0.386 (−1) = 12.400
1980	13.2	0	0	0	12.786 + 0.386 (0) = 12.786
1981	12.6	1	12.6	1	12.786 + 0.386 (1) = 13.172
1982	14.2	2	28.4	4	12.786 + 0.386 (2) = 13.558
1983	13.7	3	41.1	9	12.786 + 0.386 (3) = 13.944
Total	89.5	0	10.8	28	

The least squares trend line is $\hat{y} = a + bx$

Since $\Sigma x = 0$, the least squares estimates a and b are

$$a = \frac{\Sigma y}{n} = \frac{89.5}{7} = 12.786$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{10.8}{28} = 0.386$$

The best fitted line is

$$\hat{y} = 12.786 + 0.386\,x \qquad \text{with origin at 1980}$$

For different values of x, the trend values are obtained as shown in the table.

***Example* 16.8** *The consumer price index for medical care (medical cost) are given in the following table for the years* 1980 *to* 1987. *The base period* 1979 *is assigned the value* 100 *which actually means* 100 %.

Year	1980	1981	1982	1983	1984	1985	1986	1987
Production (000 tons)	106.0	111.1	117.2	121.3	125.2	128.0	132.6	138.0

Find a least squares linear trend,

(*i*) *by taking the origin at the middle of the time period with unit of measurement as* 1 *year*

(*ii*) *with unit of measurement as* 1/2 *year.*

Compute the trend values in both cases.

Solution. (*i*) We have, $\bar{t} = (1980 + 1987)/2 = 1983.5$. Taking the origin at the middle of the years 1983 and 1984 (*i. e.*, January 1, 1984), with unit of measurement as 1 year, we have $x = t - \bar{t} = t - 1983.5$.

Year t	Production y	Coded year $x = t - 1983.5$	xy	x^2	Trend value $\hat{y} = 122.42 + 4.38\,x$
1980	106.0	−3.5	−371.00	12.25	122.42 + 4.38 (−3.5) = 107.09
1981	111.1	−2.5	−277.75	6.25	122.42 + 4.38 (−2.5) = 111.47
1982	117.2	−1.5	−175.80	2.25	122.42 + 4.38 (−1.5) = 115.85
1983	121.3	−0.5	−60.65	0.25	122.42 + 4.38 (−0.5) = 120.23
1984	125.2	0.5	62.60	0.25	122.42 + 4.38 (0.5) = 124.61
1985	128.0	1.5	192.00	2.25	122.42 + 4.38 (1.5) = 128.99
1986	132.6	2.5	331.50	6.26	122.42 + 4.38 (2.5) = 133.37
1987	138.0	3.5	483.00	12.25	122.42 + 4.38 (3.5) = 137.75
Total	979.4	0	183.9	42	

The least squares trend line is

$$\hat{y} = a + bx$$

Since $\Sigma x = 0$, the least squares estimates a and b are

$$a = \frac{\Sigma y}{n} = \frac{979.4}{8} = 122.42$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{183.9}{42} = 4.38$$

The best fitted line is

$\hat{y} = 122.42 + 4.38\,x$ with origin at middle of the years 1983 and 1984 and unit of measurement as 1 year

For different values of x, the trend values are obtained as shown in the table.

(ii) We have, $\bar{t} = (1980 + 1987)/2 = 1983.5$. Taking the origin at the middle of the years 1983 and 1984 (*i. e.*, January 1, 1984), with unit of measurement as 1/2 year, we have

$$x = \frac{t - \bar{t}}{1/2} = \frac{t - 1983.5}{1/2}$$

Year t	Production y	Coded year $x = \frac{t - 1983.5}{1/2}$	xy	x^2	Trend value $\hat{y} = 122.42 + 2.19\,x$
1980	106.0	−7	−742.0	49	122.42 + 2.19 (−7) = 107.09
1981	111.1	−5	−555.5	25	122.42 + 2.19 (−5) = 111.47
1982	117.2	−3	−351.6	9	122.42 + 2.19 (−3) = 115.85
1983	121.3	−1	−121.3	1	122.42 + 2.19 (−1) = 120.23
1984	125.2	1	125.2	1	122.42 + 2.19 (1) = 124.61
1985	128.0	3	384.0	9	122.42 + 2.19 (3) = 128.99
1986	132.6	5	663.0	25	122.42 + 2.19 (5) = 133.37
1987	138.0	7	966.0	49	122.42 + 2.19 (7) = 137.75
Total	979.4	0	367.8	168	

The least squares trend line is $\hat{y} = a + b\,x$

Since $\Sigma x = 0$, the least squares estimates a and b are

$$a = \frac{\Sigma y}{n} = \frac{979.4}{8} = 122.42$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{367.8}{168} = 2.19$$

The best fitted line is

$\hat{y} = 122.42 + 2.19\,x$ with origin at middle of the years 1983 and 1984 and unit of measurement as 1/2 year

For different values of x, the trend values are obtained as shown in the table.

***Example* 16.9** *The consumer price index numbers y for medical care (medical cost) were given for the years* 1980 — 1987. *The base period* 1979 *was assigned the value* 100. *The least squares linear trend, with x measured from the middle of* 1983 *and* 1984 (*i. e., January* 1, 1984), *and unit of measurement as* 1/2 *year is*

$$\hat{y} = 122.42 + 2.19\,x$$

(*i*) Compute the trend values.

(*ii*) *Predict the consumer price index number for the year* 1988.

(*iii*) *In which year can we expect the index of medical cost to be double than that of* 1979 *assuming the present trends.*

Solution. The least squares linear trend is

$\hat{y} = 122.42 + 2.19\,x$ with origin at middle of the years 1983 and 1984 and unit of measurement as $1/2$ year

We have $\bar{t} = (1980 + 1987)/2 = 1983.5$. Then

$$x = \frac{t - \bar{t}}{1/2} = \frac{t - 1983.5}{1/2}$$

(*i*) For different values of x, the trend values are shown in the table.

Year t	Coded year $x = \dfrac{t - 1983.5}{1/2}$	Trend value $\hat{y} = 122.42 + 2.19\,x$
1980	−7	122.42 + 2.19 (−7) = 107.09
1981	−5	122.42 + 2.19 (−5) = 111.47
1982	−3	122.42 + 2.19 (−3) = 115.85
1983	−1	122.42 + 2.19 (−1) = 120.23
1984	1	122.42 + 2.19 (1) = 124.61
1985	3	122.42 + 2.19 (3) = 128.99
1986	5	122.42 + 2.19 (5) = 133.37
1987	7	122.42 + 2.19 (7) = 137.75

(*ii*) For $t = 1988$, we have $x = \dfrac{t - 1983.5}{1/2} = \dfrac{1988 - 1983.5}{1/2} = 9$

The estimated consumer price index for the year 1988 is

$\hat{y} = 122.42 + 2.19\,(9) = 142.13$

(*iii*) Price index for 1979 is 100. Expected price index for t is 200.

Now $200 = 122.42 + 2.19\,x \quad \Rightarrow \quad x = 35.4$

But $x = \dfrac{t - 1983.5}{1/2}$

$$35.4 = \frac{t - 1983.5}{1/2} \Rightarrow 17.7 = t - 1983.5 \Rightarrow t = 2001$$

Shifting of the Origin. While shifting the origin of a given trend line k units from the previous origin, we substitute $x + k$ or $x - k$ in the given trend line, for x depending upon whether the new origin is forward or backward of the previous origin and then find the trend line with new origin.

Thus in shifting the origin of a given linear trend the only change that take place is the change in the y-intercept. If we are to shift the origin k units forward, then to obtain the value of new y-intercept the previous y-intercept a is to be increased by k times the slope b and if we are to shift the origin k units backward, then to obtain the value of new y-intercept the previous y-intercept a is to be decreased by k times the slope b. That is, the value of the y-intercept of the new trend line would be the trend value at the new origin based on the previous trend line.

Thus, if with previous origin the trend line is $\hat{y} = a + bx$, then with new origin k units from the previous origin, the trend line is

$$\hat{y} = a + b(x \pm k) = (a \pm bk) + bx.$$

***Example* 16.10** *Suppose that the linear trend equation is* $\hat{y} = 110 + 1.5x$, *with origin at* 1980 *and unit of measurement for* x *is one year. Shift the origin at* 1985.

Solution. The linear trend equation is

$$\hat{y} = 110 + 1.5x \qquad \text{with origin at the year 1980}$$

For shifting the origin at 1985, replace x by $(x + 5)$

$$\hat{y} = 110 + 1.5(x + 5)$$
$$= 110 + 1.5x + 7.5$$
$$= 117.5 + 1.5x \qquad \text{with origin at the year 1985}$$

***Secular Trend — Nonlinear*:** Many times a straight line will not describe accurately the long-term movement of a time series. In such situations by a careful look at the graph of a time series we might detect some curvature and decide to fit a curve instead of a straight line.

***Second degree curve* (*Parabola*).** This curve is useful to describe the trend in a time series where change in the amount of change is constant per unit time. The equation of the quadratic (parabolic) trend is

$$\hat{y} = a + bx + cx^2$$

The method of least squares given the normal equations as

$$\sum y = na + b\sum x + c\sum x^2$$
$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

However, if $\sum x = 0 = \sum x^3$, then the usual normal equations reduce to

$$\sum y = na + c\sum x^2$$
$$\sum xy = b\sum x^2$$
$$\sum x^2 y = a\sum x^2 + c\sum x^4$$

which give the values of a, b *and* c as

$$c = \frac{n\sum x^2 y - (\sum x^2)(\sum y)}{n\sum x^4 - (\sum x^2)^2}$$

$$a = \frac{\sum y - c\sum x^2}{n}$$

$$b = \frac{\sum xy}{\sum x^2}$$

***Example* 16.11** *Given the following time series.*

Year	1931	1933	1935	1937	1939	1941	1943	1945
Price index	96	87	91	102	108	139	307	289

(*i*) *Fit a second-degree curve* (*parabola*) *taking the origin at* 1938.

(*ii*) *Find the trend values.*

(*iii*) *What would have been the equation of parabola if origin were at* 1933.

Solution. (*i*) We have $\bar{t} = (1931 + 1945)/2 = 1938$. Let $x = t - \bar{t} = t - 1938$

Year	Price index	Coded year					Trend value
t	y	$x = t - 1938$	x^2	x^4	xy	$x^2 y$	$\hat{y}$
1931	96	−7	49	2401	−672	4704	100.3
1933	87	−5	25	625	−435	2175	83.0
1935	91	−3	9	81	−273	819	81.8
1937	102	−1	1	1	−102	102	96.7
1939	108	1	1	1	108	108	127.7
1941	139	3	9	81	417	1251	174.7
1943	307	5	25	625	1535	7675	237.8
1945	289	7	49	2401	2023	14161	317.0
Sum	1219	0	168	6216	2601	30995	1219.0

The quadratic trend is

$$\hat{y} = a + bx + cx^2$$

Since $\Sigma x = 0 = \Sigma x^3$, the least squares estimates a, b and c are

$$c = \frac{n\Sigma x^2 y - (\Sigma x^2)(\Sigma y)}{n\Sigma x^4 - (\Sigma x^2)^2} = \frac{8(30995) - (168)(1219)}{8(6216) - (168)^2} = 2.01$$

$$a = \frac{\Sigma y - c\Sigma x^2}{n} = \frac{1219 - (2.01)(168)}{8} = 110.2$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{2601}{168} = 15.48$$

The best fitted curve is

$$\hat{y} = 110.2 + 15.48x + 2.01x^2 \qquad \text{with origin at the year 1938}$$

(*ii*) For different values of x, the trend values are obtained as shown in the table.

(*iii*) For shifting the origin at 1933, replace x by $(x - 5)$

$$\hat{y} = 110.2 + 15.48(x - 5) + 2.01(x - 5)^2$$
$$= 110.2 + 15.48(x - 5) + 2.01(x^2 - 10x + 25)$$
$$= 110.2 + 15.48x - 77.4 + 2.01x^2 - 20.1x + 50.25$$
$$= 83.05 - 4.62x + 2.01x^2 \qquad \text{with origin at the year 1933}$$

***Example* 16.12** *Given the following time series.*

Year	1931	1933	1935	1937	1939	1941	1943	1945
Price index	96	87	91	102	108	139	307	289

(*i*) *Fit a straight line taking the origin at* 1938.

(*ii*) *Fit a second-degree curve (parabola) taking the origin at* 1938.

(*iii*) *Which is the better fitted trend.*

Solution. (*i*) We have $\bar{t} = (1931 + 1945)/2 = 1938$. Let $x = t - \bar{t} = t - 1938$

Year t	Price index y	Coded year $x = t - 1938$	x^2	x^4	xy	$x^2 y$	y^2
1931	96	−7	49	2401	−672	7404	9216
1933	87	−5	25	625	−435	2175	7569
1935	91	−3	9	81	−273	819	8281
1937	102	−1	1	1	−102	102	10404
1939	108	1	1	1	108	108	11664
1941	139	3	9	81	417	1251	19321
1943	307	5	25	625	1535	7675	94249
1945	289	7	49	2401	2023	14161	83521
Sum	1219	0	168	6216	2601	30995	244225

The linear trend is

$$\hat{y} = a + bx$$

Since $\sum x = 0$, the least squares estimates a and b are

$$a = \frac{\sum y}{n} = \frac{1219}{8} = 152.38$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{2601}{168} = 15.48$$

The best fitted line is

$$\hat{y} = 152.38 + 15.48\,x \qquad \text{with origin at the year 1938}$$

The sum of squares of residuals is

$$\sum e^2 = \sum y^2 - a\sum y - b\sum xy$$

$$= 244225 - 152.38\,(1219) - 15.48\,(2601) = 18210.3$$

The quadratic trend is

$$\hat{y} = a + bx + cx^2$$

Since $\sum x = 0 = \sum x^3$, the least squares estimates a, b and c are

$$c = \frac{n\sum x^2 y - (\sum x^2)(\sum y)}{n\sum x^4 - (\sum x^2)^2} = \frac{8(30995) - (168)(1219)}{8(6216) - (168)^2} = 2.01$$

$$a = \frac{\sum y - c\sum x^2}{n} = \frac{1219 - (2.01)(168)}{8} = 110.2$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{2601}{168} = 15.48$$

The best fitted curve is

$$\hat{y} = 110.2 + 15.48\,x + 2.01\,x^2 \qquad \text{with origin at the year 1938}$$

The sum of squares of residuals is

$$\begin{aligned}\Sigma e^2 &= \Sigma y^2 - a\Sigma y - b\Sigma xy - c\Sigma x^2 y \\ &= 244225 - 110.2\,(1219) - 15.48\,(2601) - 2.01\,(30995) \\ &= 7327.77\end{aligned}$$

(iii) Since the sum of squares of the residuals for quadratic trend is smaller than the sum of squares of the residuals for linear trend, therefore quadratic trend is better fitting trend.

Merits.

(*i*) The method of least squares gives the most satisfactory measurement of the secular trend in a time series, when the distribution of the deviations is approximately normal.

(*ii*) The least squares estimates are unbiased estimates of the parameters.

(*iii*) The superiority of this method lies in that the computations needed to determine the linear, exponential or quadratic trend have been reduced to formulae.

Demerits.

(*i*) The method of least squares gives too much weight to extremely large deviations form the trend.

(*ii*) The least squares line is the best only for the period to which it has reference.

(*iii*) The elimination or addition for a few more time periods may change its position.

(*iv*) The only real criterion for the selection of a method of measuring trend is the judgement as to how well the trend line follows the general movement of the time series.

Uses of Secular Trend.

(*i*) The secular trend may be used either in determining how a time series has grown in the past or in making a forecast.

(*ii*) The trend line is used to adjust a series to eliminate the effect of the secular trend in order to isolate non-trend fluctuations.

Exercise 16.1

1. **(*a*)** What is meant by a time series? What are different movements that may be present in a time series.? Describe each of them carefully.

(*b*) Explain the difference between histogram and historigram.

(*c*) Describe the following terms:

(*i*) Secular trend. (*ii*) Seasonal variations

(*iii*) Cyclical fluctuations. (*iv*) Irregular movements.

2. **(*a*)** Describe various methods of measuring secular trend in a time series. Discuss the merits and demerits of the methods of smoothing the data.

(*b*) Plot the original time series to obtain a historigram.

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Value	50.0	36.5	43.0	44.5	38.9	38.1	32.6	38.7	41.7	41.1	33.8

Draw a free-hand trend of the following data on the same graph paper:

3. (*a*) What do you understand by the method of semi-averages utilized for smoothing of a time series. Give an example?

(*b*) The following table shows the property damaged by road accidents in Punjab for the years 1972 to 1982.

Year	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Property damaged	213	203	238	392	507	441	649	473	342	365	330

Using the method of semi-averages, find the linear trend.
($y' = 270.2 + 20.2\,x$ with origin at 1972)

(*c*) The following table gives the number of books (in 000's) sold at a book stall for the year 1970 to 1981.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981
Number of books(000)	15	18	17	42	38	40	25	20	20	16	19	17

Using semi-average method, find the trend line. Compute the trend values.
($y' = 32.01 - 1.47\,x$ with origin at 1970)

4. (*a*) What are moving averages? How is a time series smoothed by moving average method? Give an example.

(*b*) Draw a historigram of the following time series. Determine a trend line by a simple moving averages of 5-year from the following data:

Year	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930
Value	102	108	130	140	158	180	196	210	220	230

(127.6, 143.2, 160.8, 176.8, 192.8, 207.2)

5. (*a*) Calculate 7-day moving averages for the following record of attendances:

Week	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
1	24	50	30	48	54	55	62
2	28	52	41	42	50	41	42

Plot the given data and moving averages on the same graph.
(46.14, 46.71, 47.00, 48.57, 47.71, 47.14, 45.14, 42.29)

(*b*) The following table shows the United States average monthly production of bituminous coal in millions of short tons for the years 1981–91. Construct (*i*) 4-year moving averages (*ii*) 4-year centred moving averages

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Production	50.0	36.5	43.0	44.5	38.9	38.1	38.7	32.6	41.1	41.7	33.8

{ (*i*) 43.5, 40.7, 41.1, 40.0, 37.1, 37.6 38.6 37.3 (*ii*) 42.1, 40.9, 40.6 38.6, 37.4, 38.1 38.0}

6. (*a*) Compute 4-month centred moving averages from the following:

Month	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct
Value	23	26	28	30	31	35	37	32	34	38

(27.75, 29.88, 32.12, 33.50, 34.12, 34.88)

(*b*) Find 4-quarter centred moving averages for the following data.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1948	71	72	78	84
1949	72	69	75	79
1950	73	80	85	86

Plot the original data and the trend values on a graph.
(76.38, 76.12, 75.38, 74.38, 73.88, 75.38, 78.0, 80.12)

7. (*a*) Explain the method of least squares utilized for finding a secular trend in a time series.

(*b*) Given the following time series.

Year	1968	1969	1970	1971	1972	1973	1974	1975	1976
Value	3	4	6	8	7	7	10	13	12

Determine the linear trend using least squares method by taking the origin at the beginning period of the time period. Estimate the value for the year 1978.
($\hat{y} = 3.11 + 1.17x$ with origin at 1968; 14.78)

8. (*a*) The following time series shows the number of road accidents in Punjab for the years 1977 to 1987.

Year	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
Number of accidents	2493	2639	2669	3038	3745	4079	4683	4845	4505	4793	4728

(*i*) Use the method of least squares to fit a straight line taking the origin at the middle of the time period.

(*ii*) Find the trend values for this time series.

(*iii*) Estimate the number of road accidents in 1989.

{ (*i*) $\hat{y} = 3837.91 + 271.37x$ with origin at 1982; (*ii*) 2481.05, 2752.42, 3023.79, 3295.16, 3566.54, 3837.91, 4109.28, 4380.65, 4652.02, 4923.39, 5194.76; (*iii*) 5737.50 }

(*b*) Fit a straight line $\hat{y} = a + bx$ from the following results, for the years 1985—95 (both inclusive). Find out the trend values of y as well.

$$\Sigma x = 0, \quad \Sigma y = 438.9, \quad \Sigma x^2 = 110, \quad \Sigma xy = -84.4$$

($\hat{y} = 39.9 - 0.77x$ with origin at 1990; 43.75, 42.98, 42.21, 41.44, 40.67, 39.90, 39.13, 38.36, 37.59, 36.82, 36.05)

9. (*a*) Fit a straight line to the following data taking the origin at the middle of the time period and unit of measurement as 1/ 2 year and find the trend values:

Year	1980	1981	1982	1983	1984	1985
Production (000)	10	12	8	10	14	16

($\hat{y} = 11.66 + 0.54x$ with origin at the middle of 1982 and 1983 and unit of measurement as 1/ 2 year; 8.96, 10.04, 11.12, 12.20, 13.28, 14.36)

(*b*) Fit a straight line to following data. Plot on the same graph paper the actual and trend values.

Year	1970	1971	1972	1973	1974	1975	1976	1977
Value	12	15	18	25	20	22	26	30

($\hat{y} = 21 + 1.12\,x$ with origin at January 1, 1974; 13.16, 15.40, 17.64, 19.88, 22.12, 24.36, 26.60, 28.84)

(*c*) For the following time series, determine the trend by using the method of

(*i*) semi-average,

(*ii*) 3-year moving average,

(*iii*) least-squares for fitting a straight line.

Year	1968	1969	1970	1971	1972	1973	1974	1975	1976
Value	2	4	6	8	7	6	8	10	12

Which of the trend do you prefer, and why?

{ (*i*) $\hat{y} = 3.8 + 0.8\,x$ with origin at 1968, (*ii*) 4.0, 6.0, 7.0, 7.0, 7.0, 8.0, 10.0; (*iii*) $\hat{y} = 7 + x$ with origin at 1972; Least squares trend }

10. (*a*) Fit a second degree curve to the following time series and find the trend values.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production	23.2	31.4	39.8	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1

($\hat{y} = 76.64 + 13.0\,x + 0.3974\,x^2$ with origin at 1985 and unit of x as 1 year; 21.6, 31.0, 41.2 52.2, 64.0, 76.6, 90.0, 104.2, 119.2, 135.0, 151.6)

(*b*) Fit a quadratic curve to the following time series.

Year	1924	1927	1930	1933	1936	1939	1942
Index of coal price	187	142	133	129	136	169	279

Use your results to estimate the values of the index for 1935.

($\hat{y} = 119.9 + 11.89\,x + 11.99\,x^2$ with origin at 1933 and unit of x as 3 years; $\hat{y} = 133.16$)

11. (*a*) Fit a second degree curve to the following time series

Year	1980	1981	1982	1983	1984	1985	1986	1987
Quantum Index	100	87	96	102	139	210	289	307

($\hat{y} = 131.8 + 16.89\,x + 1.64\,x^2$ with origin at the January 1, 1984 and unit of x as $1/2$ year)

(*b*) Fit a quadratic curve (parabola) to the following data. Compute the trend values.

Year	1931	1933	1935	1937	1939	1941	1943	1945
Price index	96	87	91	102	108	139	307	289

($\hat{y} = 110.16 + 15.48\,x + 2.01\,x^2$ with origin at 1938 and unit of x as one year; 100.3, 83.0, 81.8, 96.7, 127.7, 174.7, 237.7, 317.0)

12. (*a*) The following are the annual profits in thousands of rupees in certain business:

Year	1977	1978	1979	1980	1981	1982	1983
Profit	88	101	105	91	113	120	132

(*i*) Fit a linear trend by the method of least-squares and make an estimate of the profits in 1985.

(*ii*) Fit a parabolic trend.

(*iii*) Determine which is the better fitting trend.

{ (*i*) $\hat{y} = 107.14 + 6.36\,x$ with origin at 1980 and unit of x as 1 year; $\hat{y} = 139.04$, (*ii*) $\hat{y} = 103.24 + 6.36\,x + 0.976\,x^2$ with origin at 1980 and unit of x as 1 year }

(*b*) Fit a quadratic trend from the following results, for the years 1985—95 (both inclusive).

$$\Sigma x = \Sigma x^3 = 0, \quad \Sigma x^2 = 110, \quad \Sigma x^4 = 1958,$$
$$\Sigma y = 410, \quad \Sigma xy = 601, \quad \Sigma x^2 y = 4587$$

Find out the trend values of y as well. Estimate the trend value for the year 1996.
($\hat{y} = 31.6 + 5.46\,x + 0.568\,x^2$ with origin at 1990 and unit of x as 1 year; 18.50, 18.85, 20.33, 22.95, 26.71, 31.60, 37.63, 44.79, 53.09, 62.53, 73.10; 84.81)

13. (*a*) Suppose that the linear trend equation is $\hat{y} = 50 + 2\,x$, with origin at 1983 and unit of measurement for x is one year. Shift the origin at 1980.

($\hat{y} = 44 + 2\,x$, with origin at the year 1980)

(*b*) If the linear trend in the data for the years 1960 to 1965 both inclusive with origin at the middle of 1962 and 1963 is $\hat{y} = 1306.667 + 73.428\,x$, the unit of x being one year, then determine the trend line with origin at 1960 and hence determine the trend values.

($\hat{y} = 1123.097 + 73.428\,x$; 1123.097, 1196.525, 1269.953, 1343.381, 1416.809, 1490.237)

(*c*) The parabolic trend equation for the projects of a company (in thousand rupees) is $\hat{y} = 10.4 + 0.6\,x + 0.7\,x^2$, with origin at 1980 and unit of measurement for x is one year. Shift the origin to 1975.

($\hat{y} = 24.9 - 6.4\,x + 0.7\,x^2$)

◆

Exercise 16.2
Objective Questions

1. With which particular characteristic movement of a time series would you mainly associate each of the following:

(*i*) Increased demand for foot-wears before Eid. (*S*)

(*ii*) The decline in death rate due to advancement in science. (*T*)

(*iii*) A steel strike, delaying production for a week. (*I*)

(*iv*) Rise in the prices of certain consumer goods due to tax increase in the annual budget. (*S*)

(*v*) An era of prosperity in a business. (*C*)

(*vi*) The festival sale. (*S*)

(*vii*) The production of sugar recorded for 1986, 1987, $\cdots$, 1992. (*T*)

(*viii*) The weekly statement of the sale of pens. (*S*)

(*ix*) A fire in a factory delaying production for 3 weeks. (*I*)

(*x*) An after Eid sale in a departmental store (*S*)

(*xi*) A need for increased wheat production due to a constant increase in population. (*T*)

(*xii*) The monthly rainfall in inches in a city over a 5-year period. (*S*)

(*xiii*) A recession in a business. (*C*)

(*xiv*) An increase in employment during summer months. (*S*)

(*xv*) A continually increasing demand for smaller automobiles. (*T*)

2. State whether the following statements are true or false.

(*i*) The graph of a time series is called histogram. (*false*)

(*ii*) Secular trend is a short term variation. (*false*)

(*iii*) Seasonal variations are regular in nature. (*true*)

(*iv*) Secular trend has booms and depressions. (*false*)

(*v*) Irregular variations are not regular in nature. (*true*)

(*iv*) The increase in the school fee in private schools is an irregular variation. (*false*)

(*vii*) The increase in the number of patients in the hospitals is like secular trend in a time series. (*true*)

(*viii*) The increase in the number of patients of heat stroke in summer is like secular trend in the time series. (*false*)

(*ix*) The secular trend is measured by a straight line when a time series has an upward trend. (*false*)

(*x*) The secular trend is measured by semi-averages method when trend is linear. (*true*)

(*xi*) The straight line is fitted to a time series when the movements in the time series are linear. (*true*)

(*xii*) In the measurement of secular trend by the method of least squares, the number of years must be odd. (*false*)

(*xiii*) For a least squares linear trend $\hat{y} = a + b\,x$, the b is a variable and $\hat{y}$ is the slope of the line. (*false*)

(*xiv*) Seasonal variations can be measured only when the time series contains yearly values. (*false*)

3. ***Multiple choice*** : Select a suitable answer:

(*i*) The graph of a time series is called

(*a*) histogram (*b*) polygon (*c*) straight line (*d*) historigram

(*ii*) The secular trend is measured by the method of semi-averages when:

(*a*) time series contains yearly values (*b*) trend is linear

(*c*) time series contains odd number of values (*d*) none of them

(*iii*) In the measurement of secular trend the moving averages

(*a*) give the trend in a straight line (*b*) measure the seasonal variations

(*c*) smooth out a time series (*d*) none of them

(*iv*) For a least squares linear trend $\hat{y} = a + b\,x$, the b is the:

(*a*) variable (*b*) intercept (*c*) trend (*d*) slope

(*v*) For a least squares linear trend $\hat{y} = a + b\,x$,

(*a*) $\Sigma y < \Sigma \hat{y}$ (*b*) $\Sigma \hat{y} = 0$ (*c*) $\Sigma y = \Sigma \hat{y}$ (*d*) none of them

(*vi*) For a least squares linear trend $\hat{y} = a + b\,x$, the $\Sigma(y - \hat{y})^2 = 0$ when

(*a*) all the y-values lie on the line. (*b*) all the y-values are positive.

(*c*) all the y-values lie above the line. (*d*) none of them.

{ (*i*) *d*, (*ii*) *b*, (*iii*) *c*, (*iv*) *d*, (*v*) *c*, (*vi*) *a* }

17 ORIENTATION OF COMPUTERS

17.1 INTRODUCTION TO COMPUTER

17.1.1 Computer. The *computer* is defined as an electronic device which is used to store and process data to solve different problems according to a set of instructions given to it. The word "computer" came from the word "compute" which means "to calculate".

17.1.2 Capabilities of Modern Computer. The following are the details of capabilities of a modern computer.

Speed. The speed of a computer is defined as the number of instructions processed in one second.

A computer can perform millions of instructions in one micro second. It performs one operation at a time. When a computer performs an operation, the clock of the processor generates electronic pulses at a fixed rate. It generates millions of pulses or signals in one second. The number of pulses generated in one second is called frequency. The unit of frequency is hertz (abbreviated Hz). Hertz is a measure of number of vibrations per second.

The speed of a computer is measured in megahertz (abbreviated MHz) and in gigahertz (abbreviated GHz). Modern personal computers may have the speed more than 3 GHz (1 GHz = 1024 MHz).

Data Storage. A computer can store a large amount of data. It stores the data in its memory and can retrieve it with a high speed. The ability of a computer to store the data and to retrieve it with a very high speed makes it suitable for modern data processing.

Data is defined as a combination of characters, numbers and symbols collected for a specified purpose.

Data Processing. Data processing consists of series of operations performed on the data to achieve the required results.

The main function of a computer is data processing. It includes the arithmetic and logical operations. It also includes the classification of data, arrangement of data and its transmission from one place to another. The results of data processing are called the output or the information.

Accuracy. The computers are very accurate in calculations.

A modern computer can perform millions of operations in one second without any error. The accuracy of calculations depends upon the input data and the program instructions. If the input data and the program instructions are correct, then we expect that the computer will produce accurate result.

Diligence. The computer has the ability to do work for long hours. It never tires. Working for long hours does not affect the accuracy of a computer.

17.2 HISTORY OF COMPUTER

The history of computer and calculator goes back to a very long way. For many centuries, people used their own brain-power to perform arithmetic calculations. The names of three great scientists who contributed in the invention of computer are

1. Abu Jaafar Muhammad Ibn Musa Al-Khwarizmi (780 — 850)
2. Alan Mathison Turing (1912 — 1954)
3. John von Neumann (1903 — 1957)

Blaise Pascal, a mathematician and scientist of France, developed the first mechanical adding machine called "Pascaline" in the 1642. Pascaline performed addition and subtraction. This machine was modified by Baron Gottfried Wilhelm von Leibnitz in 1671. He introduced "Multiplier Wheel" to perform all the basic arithmetic operations such as addition, subtraction, multiplication and division.

The designer of the first computer was Charles Babbage a mathematician of the United Kingdom. He designed a machine called "Analytical Engine" in 1837 Analytical Engine was the first programmable computer. It consisted of the following units.

(*i*) A storage (to store data)

(*ii*) A mill (to perform arithmetic operations)

(*iii*) A control unit (to control all operations and to coordinate the Input/ Output units).

The program (instructions) was given to the Analytical Engine with the help of punched cards.

The Americans were also experimenting to develop a computer. An American scientist working at Harvard University, developed a computer between 1937 and 1943. It was the "Harvard Mark-I"

In 1943, American scientists, J. W. Mauchly and J. P. Ekert developed an electronic computer at Moor School of Engineering, U.S.A. The electronic computer was called Electronic Numerical Integrator and Calculator (ENIAC). Manufacturing of ENIAC was started in 1943 and finally completed in 1946. ENIAC differed in only one significant way from the computer of today that its programs were stored externally on tape. This means that programs could be executed sequentially.

In 1944 John von Neumann suggested that the computer program should actually be stored electronically inside the computer. This was the final breakthrough in computer design.

17.3 TYPES OF COMPUTERS

The computers are of three types:

(*i*) Digital Computer

(*ii*) Analog Computer

(*iii*) Hybrid Computer

17.3.1 Digital Computer. A *digital computer* works with digits. It operates by counting numbers or digits and gives output in digital form. It works with only two signals, 0 and 1. The data and instructions are entered and stored in coded form of 0's and 1's.

These computers are manufactured in wide variety of sizes, speeds and capacities. The digital computers are commonly used in offices and educational institutions. Digital watches, digital thermometers, *etc.*, are the examples of digital computers.

17.3.2 Analog Computer. An *analog computer* does not operate directly with digital signals. It receives input gives output in the form of an analog signal.

The analog computers measure physical quantities to give output on a scale. The output is in the form of graph or a reading on a scale. A dial clock , thermometer and weighing machine are all examples of analog computers. The results achieved are not accurate as compared to those achieved by digital computers.

17.3.3 Hybrid Computer. A *hybrid computer* have features of both analog and digital computers.

The hybrid computers get input and give output either in analog or digital form. Modem is an example of hybrid computer.

17.4 CLASSIFICATION OF COMPUTERS

The computers are manufactured in a wide variety of sizes, speeds and capacities. In computer terminology, size refers to the amount of data a computer can handle. Generally a computer with a high processing speed is called a big computer. Depending upon their speed and memory size, the computers are classified into the following different groups

(*i*) Micro Computers (*iii*) Mainframe Computers

(*ii*) Mini Computers (*iv*) Super Computers

17.4.1 Micro Computers. The *micro computers* or *personal computers* are designed to be used by one user at a time. These are commonly used in offices, at homes and in educational institutions. These computers have processing speed of the order of millions of instructions processed per second (MIPS). The peripherals used in these systems include keyboard, monitor, character or page printer and a mouse.

The micro computers are small in size and are mainly used in accounting, database, word processing and graphics, *etc*. Laptop and notebooks are micro computers.

17.4.2 Mainframe Computers. The *mainframe computers* are very large computers. The mainframe computers have very high processing speed. These computers are used by large business organizations like banks, insurance companies, scientific research institutes and weather forecasting bureaus. The largest IBM *S*/390 mainframe, for example, can support 50,000 users while executing more than 1,600,000,000 instructions per second.

17.4.3 Mini Computers. The *mini computers* released in 1960s got their name because of their small size compared to the other computers of the day. They are smaller version of the mainframe computers. Like the mainframes, mini computers can handle much more data than personal computers. These are used for maintaining details of a large business organization, to analyse the results of experiments or to control and maintain the production activity in factory.

The mini computers have large memory and faster input/output devices. They are more expensive and have more processing speed than micro computers. The most powerful mini computer can serve the input and output needs of hundreds of users at a time. The mini computers cost anywhere from $ 18,000 to $ 500,000 and are ideal for many organizations and companies that cannot afford or do not need mainframe systems.

17.4.4 Super Computers. The *super computers* are the most powerful computers made, and physically they are some of the largest. These systems are built to process huge amounts of data, and the fastest super computers can perform more than 1 trillion calculations per second.

Some super computers such as the Cray T90 system can house thousands of processor. This speed and power make super computers ideal for handling large and highly complex problems that require extreme calculating power. These computers are used by Nuclear scientists

to create and analyze models of nuclear fission and fusion, predicting the action and the reactions of millions of atoms as they interact. These computers are also being to map the human genome, or DNA structure. The super computers can cost tens of millions of dollars and consume enough electricity to power dozens of homes.

17.5 HARDWARE AND SOFTWARE

17.5.1 Hardware. The physical parts of the computer are called *hardware*. It includes all physical devices or units that make up a computer. The examples of hardware are: CPU, monitor, mouse, keyboard, *etc*.

17.5.2 Software. The set of instructions given to the computer to solve a problem or to control the operation of the computer is called *software*. The, software is prepared in computer programming languages. The examples of *software* are: Microsoft Word, Excel, Corel Draw, Photoshop, *etc*.

17.6 HARDWARE COMPONENTS OF A PERSONAL COMPUTER

The computer itself, the hardware, has many parts, but the critical components fall into one of four categories.

1. Central Processing Unit (CPU)
2. Main Memory
3. Input/ Output Devices
4. Secondary Storage

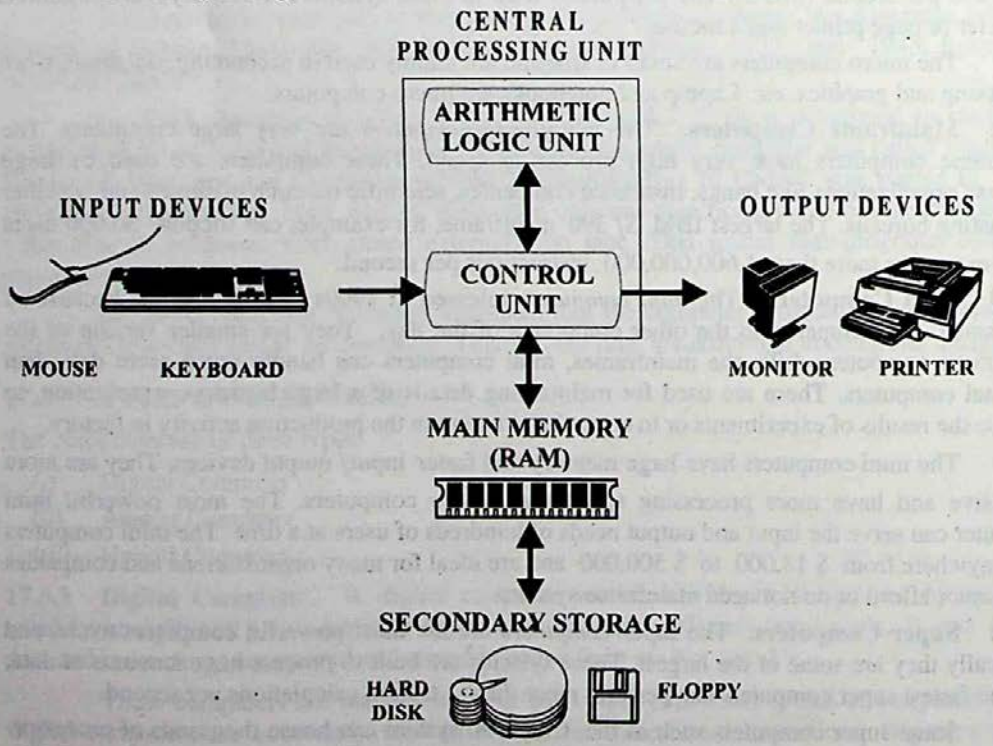


Fig. 17.1 Hardware Components of Personal Computers

17.6.1 Central Processing Unit (CPU). The *central processing unit* is the brain of the computer, the place where data is manipulated. In large computer systems, such as super computers and mainframe computers, processing tasks may be handled by multiple processing chips. (Some powerful computers systems use hundreds or even thousands of separate processing units). In average micro computer, the entire CPU is a single chip called a micro processor. The CPU has at least two basic parts:

(*i*) Control Unit

(*ii*) Arithmetic Logic Unit (ALU)

Control Unit. All the computer's resources are managed from the control unit. Think of the control unit as a traffic cop directing the flow of data through the CPU, and to the other devices The control unit is the logical hub of the computer.

The CPU's instructions for carrying out commands are built into the control unit. The instructions, or instructions set is expressed in macrocode a series of basic direction tells CPU how to execute more complex operations.

Data will have to be first transferred from the input device or secondary storage to the main memory and taken from there to the ALU for processing. Instructions on what to do with the data must be given to the ALU. Then the results have to be transferred to the main memory and from there to the output device. For these and many more such tasks we need a sort of manager. It is the control unit which takes care of all these activities.

One of the most important function of the control unit is the handling of program steps. Each basic instruction such as 'add' , 'subtract' or 'store' is in the form of code. Only the control unit understand each code and gets the instruction executed. In that process it may move data from an input device to the memory, from the memory to the ALU, from the ALU back to the memory, from memory to an output device and so on. The control unit is like the nervous system of the body and supervises all the operations of computer

***Arithmetic Logic Unit* (ALU).** The arithmetic logic unit is a part of the processor in which all arithmetic and logical operations on the data are performed.

Arithmetic section of the ALU performs basic arithmetic operations such as addition, subtraction, multiplication and division.

A logical operation is one in which data is compared. For example, whether the *first number* is greater than the *second number*, or it is less than, equal to, not equal to, grater than or equal to, *etc*. The logic section of ALU performs logical operations.

Arithmetic Operations		Logical Operations	
+	add	=, ≠	equal to, not equal to
+	subtract	>, ≯	greater that, not greater than
×	multiply	<, ≮	less than, not less than
÷	divide	≥, ≱	greater than or equal to, not greater than nor equal to
^	raised by a power	≤ ≰	less than or equal to, not less than nor equal to

The ALU includes a group of registers high speed memory locations built directly into the CPU that are used to hold the data currently being processed. For example, the control unit might load two numbers from memory into the register in the ALU. Then it might tell the ALU to

divide the two numbers (an arithmetic operation) or to see whether the numbers are equal (a logical operation)

17.6.2 Main Memory. The main memory, also called RAM (Random Access Memory) or primary storage contained in the processor unit of the computer temporarily stores data and programme instructions when the are being processed.

The main memory has many storage locations. Each memory location has a Storage Address, like a Post Box number. The computer stores or retrieves data using the address. The computer always keeps a list of data items and corresponding addresses. This is, of course, done automatically and we need not worry about it.

When the computer retrieves data from a location, it merely reads what is stored and transfer them elsewhere. It does not destroy the stored data. On the other hand, when it stores new data in a location, the previous contents in that location are lost.

The most common measurement unit for describing a computer's memory is bytes, the amount of memory it takes to store a single character, such as a letter of the alphabet or numerical.

The measurement for Computer Memory and Storage

Unit	Abbreviation	Pronounced	Approximate value (bytes)	Actual values (bytes)
Kilobyte	KB	KILL-uh-bite	1,000	1,024
Megabyte	MB	MEHG-uh-bite	1,000,000 (1 million)	1,048,576
Gigabyte	GB	GIG-uh-bite	1,000,000,000 (1 billion)	1,073,741,824
Terabyte	TB	TERR-uh-bite	1,000,000,000,000 (1 trillion)	1,099,511,627,776

Today's personal computers commonly have from 1 GB to 4 GB of memory. Some computers improve their processing efficiency by using a limited amount of high speed RAM memory between the CPU and main memory. High-speed memory used in this manner is called cache (pronounced cash) memory. Cache memory is used to store the most frequently used instructions and data. When the processor needs the next program instructions and data, it first check the cache memory. If the required instruction or data is present in cache (called a cache bit), the processor will execute faster than if the instructions or data has to retrieve from the slower main memory.

17.6.3 Input/ Output Devices. Computers would be useless if they did not provide interaction with users. They could not receive or deliver the results of their work. Input devices accept data and instructions from the user or from another computer system (such as a computer on the internet). Output devices return processed data back to the user or to another computer system.

Input Devices. Before processing unit can work, the data and programme must be entered into the computer memory, this is done by means of input devices. The most common input devices are keyboard, mouse, scanners and digital cameras.

Output Unit. There are various devices to present information in a particular manner or to deliver it at appropriate speed., *e. g.*, video display units, line printer and COM (Computer Output Microfilm).

17.6.4 Secondary Storage. A computer can function with only processing unit, memory, input and output devices. To be really useful, however, it also need a place to keep programme files and related data when it is not using them. The purpose of storage is to hold data.

It is important to understand the difference between how a computer uses main memory and how it uses secondary storage. Main memory, also called primary storage or RAM, temporarily stores programmes and data being processed. *Secondary storage*, also called auxiliary storage, stores programmes and data when they are not being processed.

The physical components or materials on which data is stored are called storage media. The hardware components that write to, and read it from, storage media are called storage devices. Two main categories of storage technology used today are magnetic storage and optical storage. Although most storage devices and media employ one technology or other, some use both.

The primary types of magnetic storage are as follows;

(*i*) Diskettes
(*ii*) Hard disks (both fixed and removable)
(*iii*) High-capacity floppy disks
(*iv*) Disk cartridges
(*v*) Magnetic tape

The primary types of optical storage are as follows;

(*i*) Compact Disk Read-Only Memory (CD-ROM)
(*ii*) Digital Versatile Disk Read-Only Memory (DVD-ROM)
(*iii*) CD-Reeordable (CD-R)
(*iv*) CD-Re Writable (CD-RW)
(*v*) Photo CD

The most common storage medium is the magnetic disk. A disk is a round, flat object that spins around its centre. Read/ write heads , which are similar to the heads of tape recorder or VCR, are used to read data from the disk or write data onto the disk. Depending on the type of disk, read/ write heads may float just above the disk's surface or may actually touch the disk.

17.7 INPUT DEVICES AND OUTPUT DEVICES

17.7.1 Input Devices. *Input devices* consist of hardware that translate data into a form the computer can process. The people readable form may be words like the ones in these sentences, but computer readable form consists of '0' and '1' or "off" and "on" electrical signals. Input hardware devices are categorized as three types

(*i*) Keyboards
(*ii*) Pointing devices
(*iii*) Source data entry devices

17.7.2. Keyboard. Keyboard is a device that converts letters, numbers and other characters into electrical signals that are machine readable by the computer processor. The keyboard may look like a type writer keyboard to which some special keys have been added. Keyboard has 3 types of keys, namely

(*i*) Alphabet keys (A, B, C, · · ·, Z, a, b, c, · · ·, z)

(*ii*) Numeric keys (1, 2, 3, 4, 5, 6, 7, 8, 9, 0)

(*iii*) Special keys (F1, F2, · · ·, F12, Alt, Ctrl, Shift, Tab, Capslock, Enter, · · ·, etc)

The standard keyboard has 101 buttons on it and now a days the keyboards with 104, 106, 110 buttons are available in the market.

17.7.3. Pointing Devices. *Pointing devices* control the position of the cursor or pointer on the screen. Pointing devices include.

(*i*) Mouse (*ii*) Light pens, *etc.*

Mouse. A mouse is a device that is rolled about on a desktop and direct a pointer on the computer's display screen the mouse has a cable that is connected to the micro computer's system unit by being plugged into a special port or socket. It has two/ three buttons, a wire or wireless. On the bottom side of the mouse is ball that translates the mouse movement into digital signals. Depending upon the software, many commands that you can execute with a mouse can also be performed through a keyboard. The following are the functions of a mouse.

(*i*) ***Point.*** Move the pointer to the desired spot on the screen, such as over a particular word or object.

(*ii*) ***Click.*** Press and quickly release, the left mouse button twice as quickly as possible.

(*iii*) ***Drag.*** the pointer to another location

(*iv*) ***Drop.*** Release the mouse button after dragging

(*v*) ***Right click.*** To make a selection using the button on the right side of the mouse which usually brings up a pop up menu

Trackball. The trackball is movable, on top of a stationery device, that is rotated with fingers or palm of the hand. Trackballs are specially suited to portable computers, which are often used in confined places such as on airplane tray, tables. Trackballs may appear on the keyboard centred below the space bar.

Joystick. A joystick is a pointing device that consist of vertical handle like a gearshift lever mounted on a base with one or two buttons

Touch pad The touch pad is a small, flat surface over which you slide your finger, using the same movement mouse.

Light Pen. The light pen is a light sensitive stylus, or pen like device connected by a wire to the computer terminal the user brings the pen to a desired point on the display screen and presses the pen button, which identifies that screen location to the computer.

17.7.4 Output Devices. The devices that are used to receive data from the CPU in binary code and convert it into readable form are called *output devices*. The output devices enable CPU to transfer information to the user and other devices.

The output device receives data from CPU in computer code and converts it into a form that a user can understand or which is readable to the other devices. For example, the binary string 01000001 from CPU represents letter '*A*' on the screen. The output is divided in two categories:

(*i*) The output that is sent to the secondary storage, *e. g.*, magnetic tape disk, *etc.* This output can be used by the CPU as input for further processing.

(*ii*) The output that can be read and used by people. This output is further divided into:

(*a*) ***Softcopy Output.*** It is the output that is temporary and is erased when the computer is switched off, *e. g.* display on the computer screen.

(*b*) ***Hardcopy Output.*** It is the output that is permanent and is always available for use, *e. g.*, print out on the paper.

Softcopy Output Devices. Softcopy output devices are used to display output on the screen. They are also called Visual Display Units (VDU). The most commonly used softcopy device are;

(*i*) Monitors

(*ii*) Pc Projectors

(*iii*) Sound Systems

Hardcopy Output Devices. The computer user usually needs output printed on the paper for permanent record. The output received from the computer on the paper is called hardcopy. The devices used to produce a hardcopy are two types

(*i*) Printer

(*ii*) Plotter

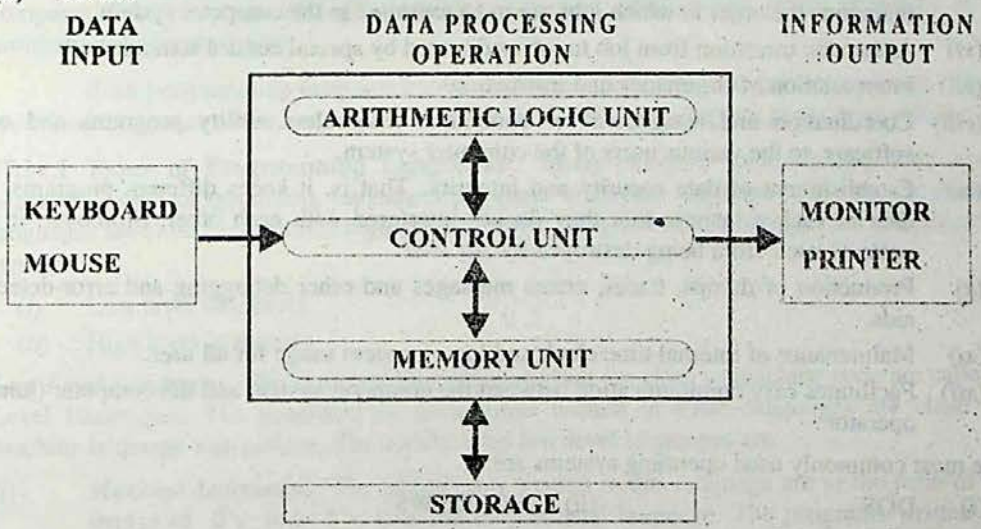


Fig. 17.2 Information Processing System

17.8 SYSTEM SOFTWARE

System software consists of all programs including the operating system that are to control the operations of the computer equipment. Some of the functions that system software perform include: starting up the computer; loading, executing, and storing application programs; storing and retrieving files; and performing a variety of functions such as formatting disks, sorting data files, and translating programs instructions into machine languages. System software can be classified into three major categories; operating systems, utilities, and language translators.

17.9 OPERATING SYSTEM

An operating system (OS) is an integrated set of programs that is used to manage the various hardware resources of computer system. Its prime objective is to improve the performance and efficiency of a computer system and increase facility, the ease with which a

system can be used. Each time a computer is turned on, or restarted the operating system is loaded into the computer and stored in the computer's main memory.

17.9.1 Functions of Operating Systems. The following are the functions of operating systems.

(*i*) Processor management, that is, assignment of processors to different tasks being performed by the computer system

(*ii*) Memory management, that is, allocation of main memory and other storage areas to system programs as well as user programs and data.

(*iii*) Input/ Output management, that is, coordination and assignment of the different input and output devices while one or more programs are being executed.

(*iv*) File management, that is, the storage of files on various storage devices and transfer of these files from one storage device to another. It also allows all files to be easily changed and modified through the use of text editors or some other file manipulation routines.

(*v*) Establishment and enforcement of job priority system. That is, it determines and maintains the order in which jobs are to be executed in the computer system.

(*vi*) Automatic transition from job to job as directed by special control statements.

(*vii*) Interpretation of commands and instructions.

(*viii*) Coordination and assignment of compilers, assemblers, utility programs and other software to the various users of the computer system.

(*ix*) Establishment of data security and integrity. That is, it keeps different programs, and data in such a manner that they do not interfered with each other. Moreover, it also protects itself from being destroyed by any user.

(*x*) Production of dumps, traces, errors messages and other debugging and error-detecting aids.

(*xi*) Maintenance of internal time clock and log of system usage for all user.

(*xii*) Facilitates easy communication between the computer system and the computer (human) operator

The most commonly used operating systems are:

(*i*) DOS (*ii*) WINDOWS

(*iii*) OS/ 2 (*iv*) UNIX

(*v*) LINUX

17.9.2 DOS. DOS stands for *Disk Operating System*, the most widely used operating system on personal computers. Several slightly different but compatible versions of DOS exist. The two most widely used, MS-DOS and PC-DOS were both originally developed by Microsoft Corporation in 1981.

MS-DOS is the text driven user interface, that is, the user types a line of text as a command. The computer then executes the command. These commands can be used to format disk, copy, and surname, delete backup files and organize and manage files on disk. MS-DOS versions 2.0 and up incorporated a tree structured hierarchical file management scheme. In this scheme, files can be managed into groups, which are known as directories. MS-DOS version 4.0 and above added additional enhanced commands and support for network and added a user interface called DOS shell with pull down menus. A ***shell*** program usually provides a limited graphic interface and certain utility functions file maintenance

17.10 APPLICATION SOFTWARE

Application software consist of programs that tell a computer how to produce information. When you think of the different ways that people uses computer in their careers or in their personal lives, your are thinking of examples of application software. Business, scientific, and educational programs are the examples of application software. The most widely used personal computer application softwares are:

(*i*) Word processing
(*ii*) Desktop publishing
(*iii*) Spreadsheet
(*iv*) Database
(*v*) Presentation graphics
(*vi*) Communications
(*vii*) Electronic mail
(*viii*) Personal information management
(*ix*) Project management

17.11 PROGRAMMING LANGUAGES

A programming language is a way of communication between the user and the computer. With the help of a programming language, programmer writes programs to solve problems with the computer.

Each programming language has its own rules for writing a computer program. The rules are called the *syntax* of the language. The process of writing a computer program is called *coding*.

17.11.1 Types of Programming Languages. Many computer programming languages are available. Some programming languages are close to human language and some programming languages are close to machine language. Therefore, programming languages are divided into two types:

(*i*) Low level languages
(*ii*) High level languages

Low Level Language. The programming language that are close to machine code are called Low Level Languages. The programs or instructions written in these languages are close to the machine language instructions. The mainly used low level languages are:

(*i*) ***Machine Language***. The instructions written in this language are in the form of binary strings of 0's and 1's. It is the fundamental language. The programs written in this language are executed directly by the computer.

(*ii*) ***Assembly Language***. It is similar to the machine language. In this language, symbolic codes are used instead of binary codes. The symbolic codes are also called mnemonic. The program written in this language is translated to machine code with the help of an assembler. This language is also known as *symbolic language*.

High Level Language. The programming languages that are close to human languages are called high level languages. The programs or instructions written in high level languages are close to English language. Each high level language has its own rules (syntax) and character set. Some of the commonly used high level languages are:

ALGOL: Algol stands for ALGOrithmic Language.
BASIC: Basic stands for Beginners All-purposes Symbolic Instruction Code.
COBOL: Cobol stands for Common Business Oriented Language.

PASCAL: This language is named in the honour of French mathematician Pascal, who invented the first mechanical calculator.

FORTRAN: Stands for FORmula TRANslation.

C: It is a general purpose language. It is widely used language in scientific and all other fields.

17.12 LANGUAGE PROCESSORS AND TRANSLATORS

The program that converts a source program, written in the programming, into the machine code, *i. e.*, in the form of strings of 0's and 1's is called language processor or translator. There are three types of language processors or translators:

(*i*) Assembler

(*ii*) Interpreter

(*iii*) Compiler

17.12.1 Assembler. An assembler translates a program written in an assembly language into machine code.

17.12.2 Interpreter. The language processors that execute a source program by translating and executing one instruction at a time are called interpreters.

17.12.3 Compiler. A compiler is a translator that coverts a program written in a high-level language.

17.13 BASIC IDEA OF WRITING AND RUNNING A COMPUTER PROGRAM

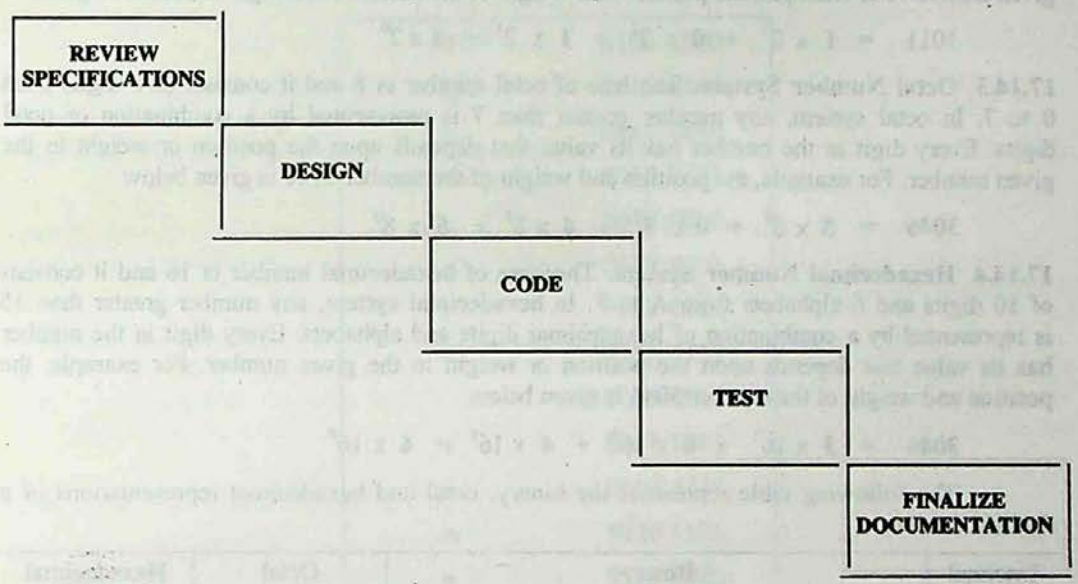
17.13.1 Computer Program. The *computer program* is a detailed set of instructions that directs a computer to perform the tasks necessary to process data into information. These instructions usually written by computer programmer, can be coded (written) in a variety of programming languages. A computer program is also known as software.

17.13.2 Computer Program Development. The program development is a process of producing one or more programs to perform specific tasks on a computer. The process of program development has evolved into a series of five steps most experts agree, should take place when any program is developed

1. ***Review specification.*** The programmer reviews the specification created by system analyst during the system design phase.
2. ***Design.*** The programmer determines and documents the specific action the computer will take to accomplish the desired tasks.

3. ***Code.*** The programmer writes the actual program instructions.
4. ***Test.*** The written programs are tested to make sure they perform as intended.
5. ***Finalize documentation.*** Throughout the program development process, the programmer documents, or writes, explanatory information about program steps 1 through 4 is brought together and organized.

Five Steps of Program Development



17.14 NUMBER SYSTEM

A set of digits, symbols and rules used to express quantities for counting, comparing amount, performing calculations, making measurements, representing values, *etc.* is called number system. A number system is named after the base of the system. The total number of digits in a number system is called its base. The most commonly used number systems are:

1. Decimal Number System.
2. Binary Number System.
3. Octal Number System.
4. Hexadecimal Number System.

The most common number system is the decimal number system. It is used in normal every day life. High level computer language nowadays use only decimal number system. Earlier programming languages required writing of long strings of numeric digits. Different number systems were used as shortcut for writing these strings. These number systems are no longer in use. However, their knowledge is necessary for understanding data representation inside the computer.

17.14.1 Decimal Number System. The base of decimal number is 10 and it consists of 10 digits from 0 to 9. In decimal system, any number greater than 9 is represented by a combination of decimal digits. Every digit in the number has its value that depends upon the

position or weight in the given number. For example, the position and weight of the number 3046 is given below

$$3046 = 3 \times 10^3 + 0 \times 10^2 + 4 \times 10^1 + 6 \times 10^0$$

17.14.2 Binary Number System. The base of binary number is 2 and it consists of 2 digits 0 and 1. In binary system, any number greater than 1 is represented by a combination of binary digits. Every digit in the number has its value that depends upon the position or weight in the given number. For example, the position and weight of the number 1011 is given below

$$1011 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$$

17.14.3 Octal Number System. The base of octal number is 8 and it consists of 8 digits from 0 to 7. In octal system, any number greater than 7 is represented by a combination of octal digits. Every digit in the number has its value that depends upon the position or weight in the given number. For example, the position and weight of the number 3046 is given below

$$3046 = 3 \times 8^3 + 0 \times 8^2 + 4 \times 8^1 + 6 \times 8^0$$

17.14.4 Hexadecimal Number System. The base of hexadecimal number is 16 and it consists of 10 digits and 6 alphabets from A to F. In hexadecimal system, any number greater than 15 is represented by a combination of hexadecimal digits and alphabets. Every digit in the number has its value that depends upon the position or weight in the given number. For example, the position and weight of the number 3046 is given below

$$3046 = 3 \times 16^3 + 0 \times 16^2 + 4 \times 16^1 + 6 \times 16^0$$

The following table represents the binary, octal and hexadecimal representations of a decimal number.

Decimal numbers	Binary representation		Octal representation	Hexadecimal representation
0	0000		00	0
1	0001		01	1
2	0010	one 2, no units	02	2
3	0011		03	3
4	0100		04	4
5	0101	one 4, no twos, one unit	05	5
6	0110		06	6
7	0111		07	7
8	1000		10	8
9	1001	one 8, no fours, no twos, one unit	11	9
10	1010		12	A
11	1011		13	B
12	1100		14	C
13	1101		15	D
14	1110		16	E
15	1111		17	F
16	10000	one 16, no eight, no fours, no twos, no unit	20	10

17.15 HOW COMPUTERS REPRESENT DATA

To a computer, every thing is a number. Numbers are numbers; letter a punctuation marks are numbers; sounds and pictures are numbers. Even computer's own instructions are numbers. When you see letters of the alphabet on a computer screen, you are seeing just one of the computer's ways of representing numbers. For example, consider this sentence: ***Here are some words.*** It may look like a string of alphabet characters to you, but to a computer it look the string of ones and zeros shown in the following table

H	0100 1000
e	0110 0101
r	0111 0010
e	0110 0101
	0010 0000
a	0110 0001
r	0111 0010
e	0110 0101
	0010 0000
s	0111 0011
o	0110 1111
m	0110 1101
e	0110 0101
	0010 0000
w	0111 0111
o	0110 1111
r	0111 0010
d	0110 0100
s	0111 0011

17.16 BINARY SYSTEM AS A FOUNDATION OF COMPUTER PROGRAMMING

In computer, however, all the data is represented by the state of the computer's electrical switches. A switch has only two possible states "on" and "off" so it can represent only two numeric values. To a computer when a switch is off, it represents a 0; when a switch is on, it represents a 1. Because there are only two values, computer are said to function in base 2, which is also known as binary number system (*bi* means " 2 " in Latin). Why we go for binary numbers instead of decimal numbers? The reasons are as fellows:

1. The first and foremost reason is that electronic and electrical components, by their very nature, operate in binary mode. Information is handled in the computer by electronic / electrical components such as transistors, semiconductors, wires, *etc.*, all of which can

only indicate two states or conditions on (1) or off (0). Transistors are either conducting (1) or non-conducting (0); magnetic materials are either magnetized (1) or non-magnetized (0) in one direction or in the opposite direction; a pulse or voltage is present (1) or not present (0) in wire. All information is represented within the computer by presence or absence of these various signals. The binary number system, which has only two digits (0 and 1), is the most suitable and conveniently used to express the two possible states.

2. The second reason is that computer circuits only have to handle two binary digits rather than ten decimal digits. The result is that the internal circuit design of computers is simplified to great extent. This ultimately results in less expensive and more reliable circuits for computers.

3. Finally, the binary number system is used because everything that can be done with base of 10 can also be done in binary.

The reason why the octal number system is used with computers is because it can represent binary values in a more compact form and because the conversation between the binary and the octal number system is very efficient.

The primary reason why the hexadecimal number systems is used with computers is because it can represent binary values in a more compact form and because the conversation between the binary and the hexadecimal number system is most efficient. An eight-digit binary number can be represented by a two-digit hexadecimal number

Exercise 17.1

1. (*a*) How are computers generally classified? What are the four major categories of computers?

   (*b*) What is CPU? Why is it called the brain of the computer?

2. (*a*) Explain the working of Arithmetic Logical Unit (ALU).

   (*a*) Explain the Control Unit.

   (*c*) What is secondary storage? How it differ from a primary storage?

3. (*a*) Describe the various input and output devices with examples.

4. (*a*) What is computers software?

   (*b*) What are the functions of a system software?

5. (*a*) What do you know about DOS?

   (*b*) What does application software do and what are its generic types?

6. (*a*) What are computer languages and their types?

   (*b*) What is an assembler?

   (*c*) What is a compiler?

7. (*a*) What is Binary Number System? Why is it used in computer?.

◆

Exercise 17.2
Objective Questions

1. Fill in the blanks.

(*i*) ———— is commonly used input device. (*keyboard*)

(*ii*) 1 MB equals ———— bytes. (1,048,576)

(*iii*) Screen output is considered as a ————. (softcopy)

(*iv*) CD-ROM is a type of ————. (Optical disk)

(*v*) ———— is a set of electronic instructions. (Software)

(*vi*) The most common type of computer memory is called ————. (RAM)

(*vii*) A high speed memory that is built into the processor is called ————. (cache memory)

(*viii*) RAM is called ———— storage. (primary)

(*xi*) Arithmetic operations are carried out by ———— unit (ALU)

(*x*) The ———— is the TV type screen that you view your programs on. (monitor)

(*xi*) The ———— allow you to type information into the computer (keyboard)

(*xii*) Keyboard, mouse, scanner are the ———— devices (input)

2. Mark off the following statements as true or false.

(*i*) 1 Kb = 1000 bytes. (true)

(*ii*) Plotter is an input device to draw the graphs of the output (false)

(*iii*) A complete computer system has two parts: hardware and software. (true)

(*iv*) The keyboard and monitor are examples of output devices. (false)

(*v*) UNIX is a application software (false)

(*vi*) The purpose of a storage device is to hold data (true)

(*vii*) Base 2 is another name for the decimal number system (false)

(*viii*) A CD-ROM is an example of a magnetic storage device (false)

(*ix*) A hard disk may also be referred to as secondary storage device. (true)

(*x*) The central processing unit (CPU) contains a Control Unit that performs arithmetic and logic operations. (false)

(*xi*) All computers work on a binary number system (true)

(*xii*) FORTRAN is a low-level language (false)