ADVANCED STATISTICS

(x_1, x_2 and x_3), there are three partial correlation coefficients. They are denoted by $r_{12.3}$, $r_{13.2}$ an $r_{23.1}$. The partial correlation coefficient $r_{12.3}$ indicates the relationship between x_1 and x_2 when th effect of x_3 is eliminated from both.

Calculation of Partial Correlation Coefficients : The formulae for calculating the partia correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ are as follows :

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Where, $r_{12.3}$ = Partial correlation between x_1 and x_2

r_{12}, r_{13} and r_{23} = Simple or zero order correlation coefficient.

Similarly, we have

$$r_{23.1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$$

Limits of Partial Correlation Coefficient : The value of $r_{12.3}$ lies betwen -1 and $+1$.

$$-1 \le r_{12.3} \le 1$$

The following examples illustrate the calculations of partial correlation coefficient.

Example 5. **Given that $r_{12} = 0.7$, $r_{13} = 0.61$, $r_{23} = 0.4$. Find the values of $r_{12.3}$, $r_{13.2}$, $r_{23.1}$.**

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Substituting the values,

$$r_{12.3} = \frac{0.7 - (0.61)(0.4)}{\sqrt{1-(0.61)^2}\sqrt{1-(0.4)^2}}$$

$$= \frac{0.456}{0.792 \times 0.916} = 0.629$$

$$r_{13.2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$$

$$= \frac{0.61 - (0.7)(0.4)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.4)^2}}$$

$$= \frac{0.61 - 0.28}{\sqrt{1-.49}\sqrt{1-.16}}$$

$$= \frac{0.33}{0.714 \times 0.916} = \frac{0.33}{0.654} = 0.505$$

$$r_{23.1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

$$= \frac{0.4 - (0.7)(0.61)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.61)^2}}$$

$$= \frac{0.4 - 0.427}{\sqrt{1-(0.49)}\sqrt{1-0.3721}} = \frac{-0.027}{0.714 \times 0.792} = -0.048$$

Example 6. **On the basis of observations made on 30 cotton plants the total correlation of yield of cotton (x_1) the number of balls *i.e.*, seed vessels (x_2) and height (x_3) are found to be:**

$$r_{12} = 0.8,\ r_{13} = 0.65,\ r_{23} = 0.7$$

Compute the partial correlation between yield of cotton and number of balls, eliminating the effect of height.

Solution. We have to find the partial correlation between yield of cotton (x_1) and the number of balls (x_2), eliminating the effect of height (x_3) *i.e.*, we have to find $r_{12.3}$.

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Substituting the given values,

$$r_{12.3} = \frac{0.8 - (0.65)(0.7)}{\sqrt{1-(.65)^2}\sqrt{1-(0.7)^2}}$$

$$= \frac{0.8 - 0.455}{\sqrt{1-0.4225}\sqrt{1-0.49}}$$

$$= \frac{0.345}{0.76 \times 0.714} = \frac{0.345}{0.543} = 0.635$$

Example 7. **For a large group of students x_1 = Score in theory, x_2 = Score in method, x_3 = Score in field work. The following results were found :**

$$r_{12} = 0.69,\ r_{13} = 0.45,\ r_{23} = 0.58$$

Determine the partial correlation coefficient between score in field work and score in theory keeping the score in method constant and interpret the result.

Solution. We have to find partial correlation coefficient between score in field (x_3) and score in theory (x_1) keeping the scores in method constant *i.e.*, we have to find $r_{31.2}$.

$$r_{31.2} = \frac{r_{31} - r_{32} \cdot r_{12}}{\sqrt{1-r_{32}^2}\sqrt{1-r_{12}^2}}$$

$$= \frac{0.45 - (0.58)(0.69)}{\sqrt{1-(0.58)^2}\sqrt{1-(0.69)^2}}$$

$$= \frac{0.45 - 0.4002}{\sqrt{1-0.3364}\sqrt{1-0.4761}}$$

Example 1. **Calculate $R_{1.23}$, $R_{3.12}$ and $R_{2.13}$ for the following data:**
$r_{12} = 0.60,\ r_{13} = 0.70,\ r_{23} = 0.65$

Solution. Given, $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$

(i) $$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2\,(0.6)\,(0.7)\,(0.65)}{1 - (0.65)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}}$$

$$= \sqrt{0.526} = 0.725$$

(ii) $$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21} \cdot r_{23} \cdot r_{13}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.65)^2 - 2\,(0.6)\,(0.65)\,(0.7)}{1 - (0.70)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{1 - 0.49}}$$

$$= \sqrt{0.4638} = 0.6809$$

(iii) $$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31} \cdot r_{32} \cdot r_{12}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.70)^2 + (0.65)^2 - 2\,(0.70)\,(0.65)\,(0.60)}{1 - (0.60)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}} = \sqrt{0.5726} = 0.756.$$

Example 2. ***For a large group of students x_1 = Score in Economics, x_2 = Score in Maths, x_3 = Score in Statistics, $r_{12} = 0.69$, $r_{13} = 0.45$, $r_{23} = 0.58$. Determine the coefficient of multiple correlation $R_{3.12}$.***

Solution. $$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31} \cdot r_{32} \cdot r_{12}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.45)^2 + (0.58)^2 - 2\,(0.45)\,(0.58)\,(0.69)}{1 - (0.69)^2}}$$

$$= \sqrt{\frac{0.2025 + 0.3364 - 0.3601}{1 - 0.4761}} = \sqrt{0.3412} = 0.584$$

Example 3. **The following zero order correlation coefficient are given:**

$$r_{12} = 0.98,\ r_{13} = 0.44 \text{ and } r_{23} = 0.54$$

Calculate multiple correlation coefficient treating the first variable as dependent and second and third variables as independent.

Solution. We have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variable as independent *i.e.*, we have to find $R_{1.23}$.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

Substituting the given values,

$$R_{1.23} = \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2\,(0.98)\,(0.44)\,(0.54)}{1 - (0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4657}{1 - 0.2916}} = \sqrt{\frac{.6883}{.7084}}$$

$$= \sqrt{0.9716} = 0.985$$

Example 4. **If $R_{1.23} = 1$, prove that $R_{2.13} = 1$.**

Solution.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

and $$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21} \cdot r_{23} \cdot r_{13}}{1 - r_{23}^2}}$$

putting $R_{1.23} = 1$ and squaring both sides,

$$1 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}$$

$$\Rightarrow \quad r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23} = 1 - r_{23}^2$$

$$\Rightarrow \quad r_{12}^2 + r_{23}^2 - 2r_{12} \cdot r_{13} \cdot r_{23} = 1 - r_{13}^2$$

$$\Rightarrow \quad \frac{r_{12}^2 + r_{23}^2 - 2r_{12} \cdot r_{23} \cdot r_{23}}{1 - r_{13}^2} = 1$$

$$\Rightarrow \quad R_{2.13}^2 = 1 \text{ or } R_{2.13} = 1$$

Since, the coefficient of multiple correlation is considered non-negative.

(2) Partial Correlation

Partial Correlation is the simple correlation between two variables after eliminating the influence of the third variable from them. For example, if we measure the relationship between yield of wheat (x_1) and the amount of fertiliser (x_2), eliminating the effect of climate (x_3) from both (having the same climate), then it is called the problem of partial correlation. For three variables

1

Partial and Multiple Correlation and Regression

INTRODUCTION

Partial and multiple correlation and regression are extension of the technique of simple correlation and regression under which we study the interrelationship between three or more variables.

(1) Multiple Correlation

Multiple correlation is the study of the relationship among three or more variables. Multiple correlation measures the combined influence of two or more independent variables on a single dependent variable. For example, if we study the combined influence of amount of fertiliser (x_2) and rainfall (x_3) on the yield of wheat (x_1), then it is called the problem of multiple correlation. We shall denote the multiple correlation coefficient between x_1, the dependent variables x_2 and x_3 independent variables by $R_{1.23}$. Similarly, we shall denote the other multiple correlation coefficients by $R_{2.13}$ and $R_{3.12}$.

Calculation of Coefficient of Multiple Correlation: The formulae for calculating the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$ are as follows :

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

Where, $R_{1.23}$ = Multiple correlation coefficient

r_{12}, r_{13}, r_{23} = Simple or zero order correlation coefficient.

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21} \cdot r_{23} \cdot r_{13}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31} \cdot r_{32} \cdot r_{12}}{1 - r_{12}^2}}$$

Limits of Multiple Correlation Coefficients: The value of multiple correlation coefficient ($R_{1.23}$) lies between 0 and 1. It can never be negative.

$$0 \leq R_{1.23} \leq 1$$

The following examples illustrate the calculations of multiple correlation coefficients:

$$= \frac{0.0498}{\sqrt{0.6636}\sqrt{0.5239}}$$

$$= \frac{0.0498}{0.81 \times 0.72} = \frac{0.0498}{0.5832} = 0.085$$

Thus, there is low degree of correlation between score in field work and score in theory.

Example 8. **Is it possible to have the following set of experimental data:**

$$r_{12} = 0.6,\ r_{23} = 0.8,\ r_{31} = -0.5.$$

Solution. In order to see whether there is inconsistency in the given data, we should calculate $r_{12.3}$. If the value of $r_{12.3}$ exceeds one, there is inconsistency, otherwise not.

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the given values,

$$= \frac{0.6 - (-0.5)(0.8)}{\sqrt{1 - (0.5)^2}\sqrt{1 - (0.8)^2}}$$

$$= \frac{0.6 + 0.4}{\sqrt{1 - 0.25}\sqrt{1 - 0.64}}$$

$$= \frac{1}{\sqrt{075}\sqrt{0.36}} = \frac{1}{0.866 \times 0.6} = \frac{1}{0.52} = 1.92$$

Since, the value of $r_{12.3}$ is greater than one, there is some inconsistency in the given data.

Aliter : We can also check the inconsistency in the data by calculating $R_{1.23}$. If the value of $R_{1.23}$ exceeds 1, there is some inconsistency otherwise not

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (-0.5)^2 - 2(0.6)(-0.5)(0.8)}{1 - (0.8)^2}}$$

$$= \sqrt{\frac{0.36 + 0.25 + 0.48}{1 - .64}} = \sqrt{\frac{1.09}{0.36}} = \sqrt{3.0277} = 1.74$$

Since, the value of $R_{1.23}$ is greater than one, there is some inconsistency in the data.

Example 9. ***Suppose a computer has found, for a given set of values of x_1, x_2 and x_3 : $r_{12} = 0.96$, $r_{13} = 0.36$ and $r_{23} = 0.78$.***

Explain whether these computations may be said to be free from errors.

olution. *For determining whether the given computed values are free from errors or not, we compute the value of $r_{12.3}$. If $r_{12.3}$ comes out to be greater than one, the computed values cannot be regarded as free from errors.*

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Substituting the given values,

$$r_{12.3} = \frac{0.96 - (0.36)(0.78)}{\sqrt{1-(0.36)^2}\sqrt{1-(078)^2}}$$

$$= \frac{0.96 - 0.2808}{\sqrt{0.8704}\sqrt{0.3916}} = \frac{0.6792}{0.9329 \times 0.6258} = \frac{0.6792}{0.5838} = 1.163$$

Since, $r_{12.3}$ is greater than one, the given computed values do contain some errors.

Relationship between Simple, Partial and Multiple Correlation Coefficients

There exists relationship between simple, partial and multiple correlation coefficients which is clear from the following equation:

(i) $1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)$

(ii) $1 - R_{2.13}^2 = (1 - r_{21}^2)(1 - r_{23.1}^2)$ and

(iii) $1 - R_{3.12}^2 = (1 - r_{31}^2)(1 - r_{32.1}^2)$

Example 10. **In a trivariate distribution, $r_{12} = 0.60$, $r_{13} = 0.70$, $r_{23} = 0.65$, find $R_{1.23}^2$ from r_{12} and $r_{13.2}$.**

Solution. Given : $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$

Multiple, Simple and Partial Correlation coefficients are related as :

$$R_{1.23}^2 = 1 - (1 - r_{12}^2)(1 - r_{13.2}^2)$$

$$r_{13.2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}} = \frac{0.70 - 0.60 \times 0.65}{\sqrt{1-(0.60)^2}\sqrt{1-(0.65)^2}}$$

$$= \frac{0.70 - 0.39}{0.8 \times 0.760} = 0.509$$

$\therefore$ $r_{13.2}^2 = 0.259, r_{12}^2 = 0.36$

Substituting values r_{12}^2 and $r_{13.2}^2$ for $R_{1.23}^2$, we have

$$R_{1.23}^2 = 1 - (1 - 0.36)(1 - 0.259)$$

$$= 1\,(0.64)(0.74) = 0.526\,.$$

MISCELLANEOUS SOLVED EXAMPLES

Exmaple 11. **x_1, x_2 and x_3 are measured from their means with :**

$$N = 10,\ \Sigma x_1^2 = 90,\ \Sigma x_2^2 = 160,\ \Sigma x_3^2 = 40$$

$$\Sigma x_1 x_2 = 60,\ \Sigma x_2 x_3 = 60,\ \Sigma x_3 x_1 = 40$$

Calculate $r_{12.3}$ and $R_{2.31}$.

Solution.

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2}\sqrt{\Sigma x_2^2}} = \frac{60}{\sqrt{90 \times 160}} = \frac{60}{120} = 0.5$$

Since, $r_{12.3}$ is greater than one, the given computations of r_{12}, r_{13} and r_{23} are not consistent.

Example 17. **If $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$, find the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.**

Solution. Given : $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{.77 - .72 \times .52}{\sqrt{1 - (.72)^2}\sqrt{1 - (.52)^2}}$$

$$= \frac{.77 - .37}{\sqrt{1 - .5184}\sqrt{1 - .2704}}$$

$$= \frac{.40}{\sqrt{.4816} \times \sqrt{.7296}} = \frac{.4}{.593} = 0.6745$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(.77)^2 + (.72)^2 - 2(.77)(.72)(.52)}{1 - (.52)^2}}$$

$$= \sqrt{\frac{.5929 + .5184 - .5766}{1 - .2704}} = \sqrt{\frac{.5347}{.7296}} = 0.856 .$$

Example 18. **If $r_{12} = 0.60$, $r_{13} = 0.70$, $r_{23} = 0.65$, find partial correlation between x_1 and x_2 and multiple correlation between x_1 dependent on x_2 and x_3.**

Solution. Given : $r_{12} = 0.60$, $r_{13} = 0.70$, $r_{23} = 0.65$

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.6 - 0.7 \times 0.65}{\sqrt{1 - (0.7)^2}\sqrt{1 - (0.65)^2}}$$

$$= \frac{0.6 - 0.455}{\sqrt{0.51 \times 0.5775}} = \frac{0.145}{0.543} = 0.2670$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(.6)^2 + (.7)^2 - 2(.6)(.7)(.65)}{1 - (.65)^2}} = \sqrt{\frac{.304}{.5775}} = 0.726 .$$

Example 19. **Following table shows the correlation matrix of three variables x_1 (Height), x_2 (Weight) and x_3 (Diameter of Chest) of 10 randomly selected players :**

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1.0000 & 0.8630 & 0.6480 \\ & 1.0000 & 0.7090 \\ & & 1.0000 \end{bmatrix} \end{array}$$

Calculate $r_{12.3}$ and $R_{1.23}$.

Solution. Given : $r_{12} = 0.863, r_{13} = 0.648, r_{23} = 0.709$

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{.863 - .648 \times .709}{\sqrt{1 - (.648)^2}\sqrt{1 - (.709)^2}}$$

$$= \frac{.863 - .4594}{\sqrt{.580} \times \sqrt{.497}} = \frac{.4036}{\sqrt{.2883}} = \frac{.4036}{.537} = 0.752$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(.863)^2 + (.648)^2 - 2(.863)(.648)(.709)}{1 - (.709)^2}}$$

$$= \sqrt{\frac{.745 + .42 - .793}{.497}} = 0.865$$

EXERCISE – 1

1. In a trivariate distribution, it is found that
$$r_{12} = 0.41, r_{13} = 0.71, r_{23} = 0.5$$
Find the value of $r_{23.1}$ and $r_{13.2}$. **[Ans. $r_{23.1} = 0.325, r_{13.2} = 0.639$]**
2. If $r_{12} = 0.7, r_{13} = 0.61$ and $r_{23} = 0.4$, find the value of $r_{12.3}, r_{13.2}$ and $r_{23.1}$. **[Ans. $r_{12.3} = 0.629, r_{13.2} = 0.505, r_{23.1} = -0.048$]**
3. Is it possible to have the following experimental data:
$$r_{12} = 0.6, r_{23} = 0.8, r_{31} = -0.5$$ **[Ans. $r_{12.3} = 1.92$, Inconsistency]**
4. In a trivariate distribution, $r_{23} = .2, r_{13} = .5, r_{12} = .6$. Compute $r_{12.3}$ and $R_{1.23}$. **[Ans. $r_{12.3} = 0.47, R_{1.23} = 0.714$]**
5. Suppose a computer has found for a given set of values of x_1, x_2, x_3 : $r_{12} = 0.91, r_{13} = 0.33$ and $r_{23} = 0.81$. Explain whether these computations may be said to be free from errors. **[Ans. $r_{12.3} = 1.161$; Not free from errors]**
6. The following zero order correlation coefficients are given:
$$r_{12} = 0.98, r_{13} = 0.44, r_{23} = 0.54$$

Example 13. **The linear correlation coefficient between x_1 (Yield), x_2 (Irrigation) and x_3 (Fertiliser) are as follows :**

$$r_{12} = 0.81,\ r_{13} = 0.90,\ r_{23} = 0.65$$

Calculate the partial correlation coefficient of:

(i) yield with irrigation

(ii) yield with fertiliser.

Solution. (i) We have to find $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

Substituting the given values,

$$r_{12.3} = \frac{(0.81) - (0.90)\,(0.65)}{\sqrt{1 - (0.90)^2}\ \sqrt{1 - (0.65)^2}}$$

$$= \frac{0.81 - 0.585}{0.4358 \times 0.7599} = \frac{0.225}{0.3311} = 0.679$$

(ii) We have to find $r_{13.2}$

$$r_{13.2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{32}^2}}$$

$$= \frac{(0.90) - (0.81)\,(0.65)}{\sqrt{1 - (0.81)^2}\ \sqrt{1 - (0.65)^2}}$$

$$= \frac{0.3735}{\sqrt{0.3439}\ \sqrt{0.5775}}$$

$$= \frac{0.3735}{0.5864 \times 0.7599} = \frac{0.3735}{0.4456} = 0.838$$

Example 14. ***Given the following zero order correlation coefficient, find (i) partial correlation coefficient between x_2 and x_3 and (ii) multiple correlation taking x_1 as dependent on x_2 and x_3.***

$$r_{12} = 0.98,\ r_{13} = 0.44,\ r_{23} = 0.54$$

Solution. ***(i)***

$$r_{23.1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{1 - r_{21}^2}\ \sqrt{1 - r_{31}^2}}$$

$$= \frac{0.54 - (0.98)\,(0.44)}{\sqrt{1 - (0.98)^2}\ \sqrt{1 - (0.44)^2}}$$

$$= \frac{0.54 - 0.4312}{\sqrt{1 - 0.9604}\ \sqrt{1 - 0.1936}}$$

$$= \frac{0.1088}{\sqrt{0.0396}\ \sqrt{0.8064}} = \frac{0.1088}{0.1786} = 0.6091$$

(ii) $$R_{1.23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$

$$=\sqrt{\frac{(0.98)^2+(0.44)^2-2(0.98)(0.44)(0.54)}{1-(0.54)^2}}$$

$$=\sqrt{\frac{0.9604+0.1936-0.4656}{1-(0.2916)}}=\sqrt{\frac{0.6884}{0.7084}}$$

$$=\sqrt{0.9717}=0.985.$$

Example 15. **Is it possible to get the following from a set of experimental data:**

(i) $r_{23}=0.8,\ r_{31}=0.5,\ r_{12}=0.6$

(ii) $r_{23}=0.7,\ r_{31}=-0.4,\ r_{12}=0.6$

Solution. (i) In order to see whether there is any inconsistency, we should calculate $r_{12.3}$, If its value exceed one, there is inconsistency, otherwise not.

$$r_{12.3}=\frac{r_{12}-r_{13}\cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$=\frac{0.6-(0.5)(0.8)}{\sqrt{1-(0.5)^2}\sqrt{1-(0.8)^2}}$$

$$=\frac{0.20}{\sqrt{0.75}\sqrt{0.36}}=\frac{0.20}{0.52}=0.384$$

Since, the value of $r_{12.3}$ is less than one, the data is consistent.

(ii) $$r_{12.3}=\frac{r_{12}-r_{13}\cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}=\frac{(0.6)-(-0.4)(0.7)}{\sqrt{1-(-0.4)^2}\sqrt{1-(0.7)^2}}$$

$$=\frac{0.6+0.28}{\sqrt{0.84}\sqrt{0.51}}=\frac{0.88}{0.655}=1.344$$

Since, the value or $r_{12.3}$ is greater than 1 there is some inconsistency in the given data.

Example 16. **Test the consistency of the following data : $r_{12}=0.8,\ r_{13}=0.4,\ r_{23}=-0.56$.**

Solution. For testing whether the given computations are consistent or not, we compute the value of $r_{13.2}$. If $r_{13.2}$ comes out to be greater than one, the computations cannot regarded as consistent.

$$r_{12.3}=\frac{r_{12}-r_{13}\cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$=\frac{(0.8)-(0.4)(-0.56)}{\sqrt{1-(0.4)^2}\sqrt{1-(-0.56)^2}}$$

$$=\frac{0.8+0.224}{\sqrt{0.84}\sqrt{0.6864}}=\frac{1.024}{0.7593}=1.349$$

$$r_{13} = \frac{\Sigma x_1 x_3}{\sqrt{\Sigma x_1^2 \times \Sigma x_3^2}} = \frac{40}{\sqrt{90 \times 40}} = \frac{40}{60} = 0.67$$

$$r_{23} = \frac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2 \times \Sigma x_3^2}} = \frac{60}{\sqrt{160 \times 40}} = \frac{60}{80} = 0.75$$

Now, $$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the values, we have

$$r_{12.3} = \frac{0.5 - 0.67 \times 0.75}{\sqrt{1 - (0.67)^2}\sqrt{1 - (0.75)^2}} = -\frac{0.0025}{0.4910} = -0.0051$$

$$R_{2.31} = \sqrt{\frac{r_{23}^2 + r_{21}^2 - 2r_{23} \cdot r_{21} \cdot r_{31}}{1 - r_{31}^2}}$$

$$= \sqrt{\frac{(0.75)^2 + (0.5)^2 - 2(0.75)(0.5)(0.67)}{1 - (.67)^2}}$$

$$= \sqrt{\frac{0.5625 + 0.25 - 0.5025}{0.5511}} = \sqrt{\frac{0.31}{0.5511}} = 0.75$$

Example 12. Calculate $r_{12.3}$ and $R_{1.23}$ from the following data :

X :	3	4	5	6	7	8	9
Y :	2	5	6	4	3	2	4
Z :	5	6	4	5	6	5	8

Solution.

Calculation of $r_{12.3}$ and $R_{1.23}$

X	X^2	Y	Y^2	Z	Z^2	XY	XZ	YZ
3	9	2	4	5	25	6	15	10
4	16	5	25	6	36	20	24	30
5	25	6	36	4	16	30	20	24
6	36	4	16	5	25	24	30	20
7	49	3	9	6	36	21	42	18
8	64	2	4	5	25	16	40	10
9	81	4	16	8	64	36	72	32
$N = 7$ $\Sigma X = 42$	$\Sigma X^2 = 280$	$\Sigma Y = 26$	$\Sigma Y^2 = 110$	$\Sigma Z = 39$	$\Sigma Z^2 = 227$	$\Sigma XY = 153$	$\Sigma XZ = 243$	$\Sigma YZ = 144$

$$r_{12} = \frac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{[\Sigma X^2 \cdot N - (\Sigma X)^2][\Sigma Y^2 \cdot N - (\Sigma Y)^2]}} = \frac{7 \times 153 - (42 \times 26)}{\sqrt{[280 \times 7 - (42)^2][110 \times 7 - (26)^2]}} = -0.155$$

$$r_{13} = \frac{N \cdot \Sigma XZ - \Sigma X \cdot \Sigma Z}{\sqrt{[\Sigma X^2 \cdot N - (\Sigma X)^2][\Sigma Z^2 \cdot N - (\Sigma Z)^2]}} = \frac{7 \times 243 - (42 \times 39)}{\sqrt{[280 \times 7 - (42)^2][227 \times 7 - (39)^2]}} = 0.546$$

$$r_{23} = \frac{N.\Sigma YZ - \Sigma Y.\Sigma Z}{\sqrt{[\Sigma Y^2.N-(\Sigma Y)^2][\Sigma Z^2.N-(\Sigma Z)^2]}} = \frac{144\times7-26\times39}{\sqrt{[110\times7-(26)^2][227\times7-(39)^2]}} = -0.075$$

Partial Correlation Coefficient

$$r_{12.3} = \frac{r_{12}-r_{13}.r_{23}}{\sqrt{1-r_{13}^2}.\sqrt{1-r_{23}^2}}$$

$$= \frac{-0.155-(0.546\times-0.075)}{\sqrt{1-(0.546)^2}\sqrt{1-(-0.075)^2}} = -0.1366$$

Multiple Correlation Coefficient

$$R_{1.23} = \sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}.r_{13}.r_{23}}{1-r_{23}^2}}$$

$$= \sqrt{\frac{(-0.155)^2+(0.546)^2-(2\times-0.155\times0.546\times-0.075)}{1-(-0.075)^2}}$$

$$= \sqrt{\frac{0.024+0.298-(.01269)}{1-.006}}$$

$$= \sqrt{\frac{0.1951}{0.994}}$$

$$= \sqrt{0.1962} = 0.443$$

Example 12 A. **In a trivariate distribution, $r_{12}=0.80$, $r_{23}=-0.56$, $r_{31}=-0.40$, compute $r_{23.1}$ and $R_{1.23}$.**

Solution. (i)

$$r_{23.1} = \frac{r_{23}-r_{21}.r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

$$= \frac{-0.56-(0.8).(-0.40)}{\sqrt{1-(0.8)^2}\sqrt{1-(-0.4)^2}}$$

$$= \frac{-0.56+0.32}{\sqrt{1-0.64}\sqrt{1-0.16}}$$

$$= \frac{-0.24}{\sqrt{0.36\times0.84}} = \frac{-0.24}{0.5499} = -0.436$$

(ii)

$$R_{1.23} = \sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}.r_{13}.r_{23}}{1-r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2+(-0.4)^2-2(0.8)(-0.4)(-0.56)}{1-(-0.56)^2}}$$

$$= \sqrt{\frac{0.64+0.16-0.3584}{1-0.3136}} = \sqrt{\frac{0.4416}{0.6864}} = 0.802$$

Calculate :

(i) the partial correlation coefficient between first (x_1) and third (x_3) variables; and

(ii) multiple correlation coefficient treating first variable (x_1) as dependent and second and third variable as independent. [Ans. $r_{13.2} = -0.53$, $R_{1.23} = 0.986$]

7. If $r_{12} = 0.9, r_{13} = 0.75, r_{23} = 0.7$, find the $R_{1.23}$. [Ans. $R_{1.23} = 0.916$]
8. Test the consistency of the data ::

$$r_{12} = 0.6, r_{13} = 0.5 \text{ and } r_{23} = 0.2.$$

Compute $r_{12.3}$, and $R_{1.23}$. [Ans. $r_{12.3} = 0.589$, Consistent]

9. Given the following values:

$$r_{12} = 0.6, r_{23} = r_{31} = 0.8,$$

find : $r_{23.1}$ and $R_{1.23}$. [Ans. $r_{23.1} = 0.667$, $R_{1.23} = 0.803$]

10. For a large group of students, x_1 = Score in Economics, x_2 = Score in Maths, x_3 = Score in Statistics, $r_{12} = 0.69, r_{13} = 0.45, r_{23} = 0.58$. Determine the coefficient of multiple correlation $R_{3.12}$. [Ans. $R_{3.12} = 0.98$]
11. The simple correlation coefficient between temperature (x_1), corp yield (x_2) and rainfall (x_3) are :

$r_{12} = 0.59, r_{13} = 0.46$ and $r_{23} = 0.77$. Calculate $r_{12.3}$ and $R_{1.23}$.

[Ans. $r_{12.3} = 0.416$, $R_{1.23} = .588$]

12. x_1, x_2 and x_3 are measured from their means with:

$N = 6, \Sigma x_1^2 = 90, \Sigma x_2^2 = 140, \Sigma x_3^2 = 4008$

$\Sigma x_1 x_2 = -100, \Sigma x_1 x_3 = -582, \Sigma x_2 x_3 = 720$

Calculate $r_{12.3}$ and $R_{1.23}$

[Ans. $r_{12} = -0.891, r_{13} = -.969, r_{23} = 0.961, r_{12.3} = 0.605, R_{1.23} = 0.97$]

(3) MULTIPLE REGRESSION

In multiple regression, we study three variables and we consider one variable as dependent variable and the other two as independent variables. Multiple regression analysis is used to estimate the most probable value of the dependent variable for given values of the independent variables.

Methods to obtain Multiple Regression Equations

Multiple regression equations can be worked out by two methods, which are as follows :

(1) Multiple Regression Equations using Normal Equations

(2) Multiple Regression Eqautions in terms of Simple Correlation Coefficients

Let us discuss them.

(1) Multiple Regression Equations using Normal Equations : This method is also called as **Least Square Method.** Under this method computation of regression equations is done by solving three normal equations. This method becomes clear by the following :

Multiple regression equation of X_1 on X_2 and X_3 is given by :

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Where, X_1 = Dependent variable, X_2 and X_3 = Independent variables.

$b_{12.3}$ and $b_{13.2}$ = Partial regression coefficients.

Using least square method, the values of constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations:

$$\Sigma X_1 = N.a_{1.23} + b_{12.3}\,\Sigma X_2 + b_{13.2}\,\Sigma X_3 \quad ...(1)$$

$$\Sigma X_1X_2 = a_{1.23}\,\Sigma X_2 + b_{12.3}\,\Sigma X_2^2 + b_{13.2}\,\Sigma X_2X_3 \quad ...(2)$$

$$\Sigma X_1X_3 = a_{1.23}\,\Sigma X_3 + b_{12.3}\,\Sigma X_2X_3 + b_{13.2}\,\Sigma X_3^2 \quad ...(3)$$

Similarly, the multiple regression equations of X_2 on X_1 and X_3 and X_3 ; on X_1 and X_2 and their normal equations can also be written as:

Multiple Regression Equation of X_2 on X_1 and X_3 is given by:

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$$

Three Normal Equations are :

$$\Sigma X_2 = N a_{2.13} + b_{21.3}\,\Sigma X_1 + b_{23.1}\,\Sigma X_3$$

$$\Sigma X_2X_1 = a_{2.13}\,\Sigma X_1 + b_{21.3}\Sigma X_1^2 + b_{23.1}\,\Sigma X_3X_1$$

$$\Sigma X_2X_3 = a_{2.13}\,\Sigma X_3 + b_{21.3}\,\Sigma X_1X_3 + b_{23.1}\,\Sigma X_3^2$$

Multiple Regression Equation of X_3 on X_1 and X_2 is given by:

$$X_3 = a_{3.12} + b_{31.2}\,X_1 + b_{32.1}\,X_2$$

Three Normal Equations are:

$$\Sigma X_3 = N a_{3.12} + b_{31.2}\,\Sigma X_1 + b_{32.1}\,\Sigma X_2$$

$$\Sigma X_3X_1 = a_{3.12}\,\Sigma X_1 + b_{31.2}\,\Sigma X_1^2 + b_{32.1}\,\Sigma X_2X_1$$

$$\Sigma X_3X_2 = a_{3.12}\,\Sigma X_2 + b_{31.2}\,\Sigma X_1X_2 + b_{32.1}\,\Sigma X_2^2$$

The following example illustrate the procedure of fitting multiple regression equations:

Example 1. **For the following set of data, calculate multiple regression equation of X_1 on X_2 and X_3 :**

X_1 :	4	6	7	9	13	15
X_2 :	15	12	8	6	4	3
X_3 :	30	24	20	14	10	4

Solution. The regression equation of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3}\,X_2 + b_{13.2}\,X_3$$

The three normal equations are :

$$\Sigma X_1 = N a_{1.23} + b_{12.3}\,\Sigma X_2 + b_{13.2}\,\Sigma X_3$$

$$\Sigma X_1X_2 = a_{1.23}\,\Sigma X_2 + b_{12.3}\,\Sigma X_2^2 + b_{13.2}\,\Sigma X_3X_2$$

$$\Sigma X_1X_3 = a_{1.23}\,\Sigma X_3 + b_{12.3}\,\Sigma X_2X_3 + b_{13.2}\,\Sigma X_3^2$$

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2
4	15	30	60	120	450	225	900
6	12	24	72	144	288	144	576
7	8	20	56	140	160	64	400
9	6	14	54	126	84	36	196
13	4	10	52	130	40	16	100
15	3	4	45	60	12	9	16
$\Sigma X_1 = 54$	$\Sigma X_2 = 48$	$\Sigma X_3 = 102$	$\Sigma X_1X_2 = 339$	$\Sigma X_1X_3 = 720$	$\Sigma X_2X_3 = 1034$	$\Sigma X_2^2 = 494$	$\Sigma X_3^2 = 2188$

Substituting the values in the normal equations :

$$54 = 6a_{1.23} + 48b_{12.3} + 102b_{13.2} \quad ...(i)$$

$$339 = 48a_{1.23} + 494b_{12.3} + 1034b_{13.2} \quad ...(ii)$$

$$720 = 102a_{1.33} + 1034b_{12.3} + 2188b_{13.2} \quad ...(iii)$$

Multiplying (i) by 8, we get

$$432 = 48a_{1.23} + 384b_{12.3} + 816b_{13.2} \quad ...(iv)$$

Subtracting (ii) from (iv), we get

$$-93 = 110b_{12.3} + 218b_{13.2} \quad ...(v)$$

Multiplying (i) by 17, we get

$$918 = 102a_{1.23} + 816b_{12.3} + 1734b_{13.2} \quad ...(vi)$$

Subtracting (iii) from (vi), we get

$$-198 = 218b_{12.3} + 454b_{13.2} \quad ...(vii)$$

Multiplying (v) by 109, we obtain

$$-10137 = 11990b_{12.3} + 23762b_{13.2} \quad ...(viii)$$

Multiplying (vii) by 55, we get

$$-10890 = 11990b_{12.3} + 24970b_{13.2} \quad ...(ix)$$

Subtracting (viii) from (ix), we get

$$753 = -1208b_{13.2}$$

$$\therefore \quad b_{13.2} = -\frac{753}{1208} = -0.623$$

Substituting the value of $b_{13.2}$ in equation (v), we get

$$-93 = 110b_{12.3} + 218(-0.623)$$

$$135.814 - 93 = 110b_{12.3}$$

$$\therefore \quad b_{12.3} = \frac{42.814}{110} = 0.389$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (i), we get

$$6a_{1.23} + 48(0.389) + 102(-0.623) = 54$$

$$6a_{1.23} + 18.672 - 63.546 = 54$$

$$6a_{1.23} = 54 - 18.672 + 63.546$$

$$6a_{1.23} = 98.874$$

$$\therefore \quad a_{1.23} = \frac{98.874}{6} = 16.479$$

Hence, the required equation is

$$X_1 = 16.479 + 0.389X_2 - 0.623X_3$$

EXERCISE – 2

1. *From the following data, find the least square regression of X_1, X_2 and X_3 and estimate the value of X_1 for given values of $X_2 = 16$ and $X_3 = 4$:*

X_1 :	10	5	10	4	8
X_2 :	16	13	21	10	13
X_3 :	3	6	4	5	3

[Ans. $X_1 = 4.753 + 0.502X_2 - 1.115X_3$, 8.325]

2. Compute the values of b_0, b_1 and b_2 for the equation $Y = b_0 + b_1X_1 + b_2X_2$ from the following data:

Y:	3	5	6	8	12	14
X_1:	16	10	7	4	3	2
X_2:	90	72	54	42	30	12

[Ans. $Y = 16.1067 + .426X_1 - 0.221X_2$]

3. Obtain the parameters of the multiple linear regression model: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3$ from the following data:

$$N = 6, \Sigma Y = 54, \Sigma X_2 = 48, \Sigma X_3 = 102$$
$$\Sigma YX_2 = 339, \Sigma YX_3 = 720, \Sigma X_2X_3 = 1034, \Sigma X_2^2 = 494, \Sigma X_3^2 = 2188$$

[Ans. $Y = 16.479 + 0.389X_2 - 0.623X_3$]

Short-Cut Method : When the size of the values of the variables are very large, then the above system of solving normal equations becomes a very tedious procedure. In such a case, in place of actual values, deviations from the means of the variables are used to simplify the computation procedure.

Multiple Regression Eqaution of X_1 on X_2 and X_3 in deviation form is given by :

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

or $$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{where } x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2, x_3 = X_3 - \bar{X}_3$$

The values of the partial regression coefficients ($b_{12.3}$ and $b_{13.2}$) can be obtained by solving the following two normal equations :

$$\Sigma x_1 x_2 = b_{12.3}\Sigma x_2^2 + b_{13.2}\Sigma x_3 x_2$$
$$\Sigma x_1 x_3 = b_{12.3}\Sigma x_2 x_3 + b_{13.2}\Sigma x_3^2$$

Further solved, we have

$$b_{12.3} = \frac{(\Sigma x_1 x_2)(\Sigma x_3^2) - (\Sigma x_1 x_3)(\Sigma x_2 x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}$$

$$b_{13.2} = \frac{(\Sigma x_1 x_3)(\Sigma x_2^2) - (\Sigma x_1 x_2)(\Sigma x_3 x_2)}{(\Sigma x_3^2)(\Sigma x_2^2) - (\Sigma x_3 x_2)^2}$$

Similarly, the multiple regression equation of X_2 on X_1 and X_3 ; and X_3 on X_1 and X_2 and their normal equations can also be written.

Multiple Regression Equation of X_2 on X_1 and X_3 in deviation form is given by:

$$X_2 - \bar{X}_2 = b_{21.3}(X_1 - \bar{X}_1) + b_{23.1}(X_3 - \bar{X}_3)$$

or $$x_2 = b_{21.3}x_1 + b_{23.1}x_3$$

Two normal equations are :

$$\Sigma x_2 x_1 = b_{21.3}\Sigma x_1^2 + b_{23.1}\Sigma x_1 x_3$$
$$\Sigma x_2 x_3 = b_{21.3}\Sigma x_1 x_3 + b_{23.1}\Sigma x_3^2$$

Further solved, we have

$$b_{21.3} = \frac{(\Sigma x_2 x_1)(\Sigma x_3^2) - (\Sigma x_2 x_3)(\Sigma x_1 x_3)}{(\Sigma x_1^2)(\Sigma x_3^2) - (\Sigma x_1 x_3)^2}$$

$$b_{23.1} = \frac{(\Sigma x_2 x_3)(\Sigma x_1^2) - (\Sigma x_2 x_1)(\Sigma x_3 x_1)}{(\Sigma x_3^2)(\Sigma x_1^2) - (\Sigma x_3 x_1)^2}$$

Multiple Regression Equation of X_3 on X_1 and X_2 in deviation form is given by:

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

or $$x_3 = b_{31.2}\, x_1 + b_{32.1} x_2$$

Two normal equations are :

$$\Sigma x_1 x_3 = b_{31.2}\, \Sigma x_1^2 + b_{23.1}\, \Sigma x_1 x_2$$

$$\Sigma x_2 x_3 = b_{31.2}\, \Sigma x_1 x_2 + b_{32.1}\, \Sigma x_2^2$$

Further solved, we have

$$b_{31.2} = \frac{(\Sigma x_3 x_1)(\Sigma x_2^2) - (\Sigma x_3 x_2)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$b_{32.1} = \frac{(\Sigma x_3 x_2)(\Sigma x_1^2) - (\Sigma x_3 x_1)(\Sigma x_2 x_1)}{(\Sigma x_2^2)(\Sigma x_1^2) - (\Sigma x_2 x_1)^2}$$

The following examples would clarify the method:

Example 1. **From the following data, find the least square regression of X_3 on X_1 and X_2 using actual mean method. Also estimate X_3 when $X_1 = 10$ and $X_2 = 6$.**

X_1:	3	5	6	8	12	14
X_2:	16	10	7	4	3	2
X_3:	90	72	54	42	30	12

Solution.

X_1	$x_1 = (X_1 - \bar{X}_1)$	x_1^2	X_2	$x_2 = (X_2 - \bar{X}_2)$	x_2^2	X_3	$x_3 = (X_3 - \bar{X}_3)$	x_3^2	x_1x_2	x_1x_3	x_2x_3
3	−5	25	16	+9	81	90	+40	1600	−45	−200	+360
5	−3	9	10	+3	9	72	+22	484	−9	−66	+66
6	−2	4	7	0	0	54	+4	16	0	−8	0
8	0	0	4	−3	9	42	−8	64	0	0	+24
12	+4	16	3	−4	16	30	−20	400	−16	−80	+80
14	+6	36	2	−5	25	12	−38	1444	−30	−228	+190
$\Sigma X_1 = 48$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 90$	$\Sigma X_2 = 42$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 140$	$\Sigma X_3 = 300$	$\Sigma x_3 = 0$	$\Sigma x_3^2 = 4008$	$\Sigma x_1x_2 = -100$	$\Sigma x_1x_3 = -582$	$\Sigma x_2x_3 = 720$

$$\bar{X}_1 = \frac{48}{6} = 8, \bar{X}_2 = \frac{42}{6} = 7, \bar{X}_3 = \frac{300}{6} = 50,$$

Regression Equation of X_3 on X_1 and X_2 is:

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

$$b_{31.2} = \frac{(\Sigma x_3 x_1)(\Sigma x_2^2) - (\Sigma x_3 x_2)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{(-582)(140) - (720)(-100)}{(90)(140) - (-100)^2}$$

$$= \frac{-81480+72000}{12600-10000} = \frac{-9480}{2600} = -3.646$$

$$b_{32.1} = \frac{(\Sigma x_3 x_2)(\Sigma x_1^2) - (\Sigma x_3 x_1)(\Sigma x_2 x_1)}{(\Sigma x_2^2)(\Sigma x_1^2) - (\Sigma x_2 x_1)^2}$$

$$= \frac{(720)(90) - (-582)(-100)}{(90)(140) - (-100)^2}$$

$$= \frac{64800-58200}{12600-10000} = \frac{6600}{2600} = 2.538$$

Substituting the values in the above equations, we get

$$X_3 - 50 = -3.646(X_1 - 8) + 2.538(X_2 - 7)$$

$$X_3 - 50 = -3.646X_1 + 29.168 + 2.538X_2 - 17.766$$

$$X_3 = -3.646X_1 + 2.538X_2 + 61.402$$

When $X_1 = 10$ and $X_2 = 6$, So, $X_3 = -3.646(10) + 2.538(6) + 61.402$

$$= -36.46 + 15.228 + 61.402 = 40.17 \text{ or } 40.$$

Example 3. **Given the following information (variables are measured from their respective means):**

$\Sigma x_1 x_2 = 720$, $\Sigma x_2 x_3 = -582$, $\Sigma x_1 x_3 = -100$

$\Sigma x_2^2 = 4008$, $\Sigma x_3^2 = 90$, $\Sigma x_1^2 = 140$

$\bar{X}_1 = 7$, $\bar{X}_2 = 50$, $\bar{X}_3 = 8$

Find the multiple regression equation of X_1 on X_2 and X_3. Estimate X_1 when $X_2 = 10$ and $X_3 = 95$.

Solution. Regression Equation of X_1 on X_2 and X_3 is given by :

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

$$b_{12.3} = \frac{(\Sigma x_1 x_2)(\Sigma x_3^2) - (\Sigma x_1 x_3)(\Sigma x_2 x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}$$

$$= \frac{(720)(90) - (-100)(-582)}{(4008)(90) - (-582)^2}$$

$$= \frac{64800-58200}{360720-338724} = \frac{6600}{21996} = 0.30$$

$$b_{13.2} = \frac{(\Sigma x_1 x_3)(\Sigma x_2^2) - (\Sigma x_1 x_2)(\Sigma x_3 x_2)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}$$

$$= \frac{(-100)(4008) - (720)(-582)}{(4008)(90) - (-582)^2}$$

$$= \frac{-400800+419040}{360720-338724} = \frac{18240}{21996} = 0.829 = 0.83$$

We are given : $\bar{X}_1 = 7$, $\bar{X}_2 = 50$, $\bar{X}_3 = 8$

Substituting the values in the above equation, we get

$$X_1 - 7 = 0.30(X_2 - 50) + 0.83(X_3 - 8)$$

or $$X_1 - 7 = 0.30X_2 - 15 + 0.83X_3 - 6.64$$

$\therefore X_1 = 0.30X_2 + 0.83X_3 - 14.64$ is the required equation

When $X_2 = 20$ and $X_3 = 30$

$$X_1 = 0.30(20) + 0.83(30) - 14.64 = 6 + 24.9 - 14.64 = 16.26$$

Example 4. **The following data for three variables X_1, X_2 and X_3 are given below :**

$\Sigma x_1x_2 = 218$, $\Sigma x_1x_3 = -198$, $\Sigma x_2x_3 = -93$

$\Sigma x_1^2 = 454$, $\Sigma x_2^2 = 110$, $\Sigma x_3^2 = 90$

x_1, x_2 and x_3 are measured from their means. Find the two partial regression coefficients ($b_{12.3}$ and $b_{13.2}$).

Solution.

$$b_{12.3} = \frac{(\Sigma x_1x_2)(\Sigma x_3^2) - (\Sigma x_1x_3)(\Sigma x_2x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2x_3)^2}$$

$$= \frac{(218)(90) - (-198)(-93)}{(90)(110) - (-93)^2}$$

$$= \frac{19620 - 18414}{9900 - 8649}$$

$$= \frac{1206}{1251} = 0.964$$

$$b_{13.2} = \frac{(\Sigma x_1x_3)(\Sigma x_2^2) - (\Sigma x_1x_2)(\Sigma x_3x_2)}{(\Sigma x_3^2)(\Sigma x_2^2) - (\Sigma x_3x_2)^2}$$

$$= \frac{(-198)(110) - (218)(-93)}{(90)(110) - (-93)^2}$$

$$= \frac{-21780 + 20274}{9900 - 8649} = \frac{-1506}{1251} = -1.203$$

EXERCISE – 3

1. *For the following set of data, find the multiple regression of X_1 on X_2 and X_3 using actual mean method. Also predict the value of X_1 when $X_2 = 5$ and $X_3 = 7$:*

X_1 :	12	24	32	28	
X_2 :	6	12	16	22	
X_3 :	4	6	12	18	

[Ans. $X_1 = 2.577 + 1.661X_2 + 0.0169X_3$, $X_1 = 110$]

2. *From the data given below, find the multiple regression equation of X_1 on X_2 and X_3 using actual mean method :*

X_1 :	4	6	7	9	13	15
X_2 :	15	12	8	6	4	3
X_3 :	30	24	20	14	10	4

[Ans. $X_1 = 16.479 + 0.389X_2 + 0.623X_3$]

3. From the data given below, find the multiple linear regression of X_1 on X_2 and X_3 using actual mean method :

X_1 :	18	20	17	14	21
X_2 :	38	40	25	28	44
X_3 :	20	15	5	12	18

[Ans. $X_1 = 0.5\,X_2 - 0.36\,X_3 + 5.54$]

4. Given the following information (variables are measured from their respective means):

$\Sigma x_1 x_2 = 1900$, $\Sigma x_1 x_3 = -20$, $\Sigma x_2 x_3 = -50$,
$\Sigma x_1^2 = 1350$, $\Sigma x_2^2 = 2800$, $\Sigma x_3^2 = 24$,
$\overline{X}_1 = 65$, $\overline{X}_2 = 55$, $\overline{X}_3 = 30$,

Obtain the partial regression coefficients ($b_{12.3}$ and $b_{13.2}$)
Also estimate the value of X_1 when $X_2 = 60$ and $X_3 = 25$.

[Ans. $b_{12.3} = 0.689$, $b_{13.2} = 0.603$, $X_1 = 65.43$]

5. Given the following information (variates are measured from their respective means) :

$\Sigma x_1^2 = 1350$, $\Sigma x_2^2 = 2800$, $\Sigma x_3^2 = 24$

$\Sigma x_1 x_2 = 1900$, $\Sigma x_1 x_3 = -20$, $\Sigma x_2 x_3 = -50$

Determine the regression equation of X_1 on X_2 and X_3. [Ans. $X_1 = 0.689\,X_2 + .603\,X_3$]

---

(2) Multiple Regression Equations in terms of Simple Correlation Coefficients

When the values of $\overline{X}_1$, $\overline{X}_2$ and $\overline{X}_3$, σ_1, σ_2 and σ_3 and r_{12}, r_{13} and r_{23} are given, then the multiple regression equations are expressed in the following manner :

(1) Multiple Regression Equation of X_1 on X_2 and X_3

$$X_1 - \overline{X}_1 = b_{12.3}\,(X_2 - \overline{X}_2) + b_{13.2}\,(X_3 - \overline{X}_3)$$

or $$x_1 = b_{12.3}\,x_2 + b_{13.2}\,x_3 \qquad \text{Where, } x_1 = X_1 - \overline{X}_1,\ x_2 = X_2 - \overline{X}_2,\ x_3 = X_3 - \overline{X}_3$$

The values of partial regression coefficients $b_{12.3}$ and $b_{13.2}$ are determined by using the following formulae :

$$b_{12.3} = \left[\frac{\sigma_1}{\sigma_2}\right]\cdot\left[\frac{r_{12} - r_{13}\cdot r_{23}}{1 - r_{23}^2}\right]$$

$$b_{13.2} = \left[\frac{\sigma_1}{\sigma_3}\right]\cdot\left[\frac{r_{13} - r_{12}\cdot r_{32}}{1 - r_{32}^2}\right]$$

Multiple Regression Equation of X_1 on X_2 and X_3 can also be written as :

$$x_1 = \left[\frac{\sigma_1}{\sigma_2}\right]\left[\frac{r_{12} - r_{13}\cdot r_{23}}{1 - r_{23}^2}\right]x_2 + \left[\frac{\sigma_1}{\sigma_3}\right]\left[\frac{r_{13} - r_{12}\cdot r_{32}}{1 - r_{32}^2}\right]x_3$$

or $$X_1 - \overline{X}_1 = \left[\frac{\sigma_1}{\sigma_2}\right]\left[\frac{r_{12} - r_{13}\cdot r_{23}}{1 - r_{23}^2}\right](X_2 - \overline{X}_2) + \left[\frac{\sigma_1}{\sigma_3}\right]\left[\frac{r_{13} - r_{12}\cdot r_{32}}{1 - r_{23}^2}\right](X_3 - \overline{X}_3)$$

(2) Multiple Regression Equation of X_2 on X_1 and X_3 :

$$X_2 - \overline{X}_2 = b_{21.3}\,(X_1 - \overline{X}_1) + b_{23.1}\,(X_3 - \overline{X}_3)$$

or $$x_2 = b_{21.3}\,x_1 + b_{23.1}\,x_3$$

Where,

$$b_{21.3} = \left[\frac{\sigma_2}{\sigma_1}\right]\left[\frac{r_{21} - r_{23} \cdot r_{13}}{1 - r_{13}^2}\right]$$

$$b_{23.1} = \left[\frac{\sigma_2}{\sigma_3}\right]\left[\frac{r_{23} - r_{21} \cdot r_{31}}{1 - r_{31}^2}\right]$$

Multiple Regression Equation of X_2 on X_1 and X_3 can also be written :

$$x_2 = \left[\frac{\sigma_2}{\sigma_1}\right]\left[\frac{r_{21} - r_{23} \cdot r_{13}}{1 - r_{13}^2}\right] x_1 + \left[\frac{\sigma_2}{\sigma_3}\right]\left[\frac{r_{23} - r_{21} \cdot r_{31}}{1 - r_{31}^2}\right] x_3$$

or $$X_2 - \bar{X}_1 = \left[\frac{\sigma_2}{\sigma_1}\right]\left[\frac{r_{21} - r_{23} \cdot r_{13}}{1 - r_{13}^2}\right](X_1 - \bar{X}_1) + \left[\frac{\sigma_2}{\sigma_3}\right]\left[\frac{r_{23} - r_{21} \cdot r_{31}}{1 - r_{31}^2}\right](X_3 - \bar{X}_3)$$

(3) Multiple Regression Equation of X_3 on X_1 and X_2 :

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

or $$x_3 = b_{31.2}\, x_1 + b_{32.1}\, x_2$$

Where $$b_{31.2} = \left[\frac{\sigma_3}{\sigma_1}\right]\left[\frac{r_{31} - r_{32} \cdot r_{12}}{1 - r_{12}^2}\right]$$

$$b_{32.1} = \left[\frac{\sigma_3}{\sigma_2}\right]\left[\frac{r_{32} - r_{31} \cdot r_{21}}{1 - r_{21}^2}\right]$$

Multiple regression on X_3 on X_1 and X_2 can also be written as :

$$x_3 = \left[\frac{\sigma_3}{\sigma_1}\right]\left[\frac{r_{31} - r_{32} \cdot r_{12}}{1 - r_{12}^2}\right] x_1 + \left[\frac{\sigma_3}{\sigma_2}\right]\left[\frac{r_{32} - r_{31} \cdot r_{21}}{1 - r_{21}^2}\right] x_2$$

or $$X_3 - \bar{X}_3 = \left[\frac{\sigma_3}{\sigma_1}\right]\left[\frac{r_{31} - r_{32} \cdot r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1) + \left[\frac{\sigma_3}{\sigma_2}\right]\left[\frac{r_{32} - r_{31} \cdot r_{21}}{1 - r_{21}^2}\right](X_2 - \bar{X}_2)$$

Note : $r_{12} = r_{21}, r_{23} = r_{32}, r_{13} = r_{31}$.

The following examples would clarify the procedure :

Example 5. **A teacher in mathematics wishes to determine the relationship of marks in final examination to those in two tests given during the semester. Calling X_1, X_2 and X_3, the marks of a student on Ist, 2nd and final examination respectively, he made the following computations from a total of 120 students :**

$\bar{X}_1 = 6.8$ $\quad \bar{X}_2 = 7.0$ $\quad \bar{X}_3 = 74$

$\sigma_1 = 1.0$ $\quad \sigma_2 = 0.80$ $\quad \sigma_3 = 9.0$

$r_{12} = 0.60$ $\quad r_{13} = 0.70$ $\quad r_{23} = 0.65$

(i) Find the relevant regression equation.

(ii) Estimate the final marks of two students who secured respectively 9 and 7.4 and 8 on the two tests.

Solution. The relevant least square regression equation will be X_3 on X_1 and X_2 which is given by :

$$X_3 - \overline{X}_3 = b_{31.2}(X_1 - \overline{X}_1) + b_{32.1}(X_2 - \overline{X}_2)$$

$$b_{31.2} = \frac{\sigma_3}{\sigma_1} \cdot \left[\frac{r_{31} - r_{32} \cdot r_{12}}{1 - r_{12}^2}\right]$$

$$= \frac{9}{1} \times \left[\frac{(.70) - (.65)(.60)}{1 - (.60)^2}\right]$$

$$= 9 \times \left[\frac{(.70) - (.39)}{1 - (.60)^2}\right] = 9 \times \left[\frac{.31}{.64}\right] = \frac{2.79}{.64} = 4.36$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2} \cdot \left[\frac{r_{32} - r_{31} \cdot r_{21}}{1 - r_{21}^2}\right]$$

$$= \frac{9}{.80} \times \left[\frac{(.65) - (.70)(.60)}{1 - (.60)^2}\right]$$

$$= \frac{9}{.80} \times \left[\frac{.65 - .42}{0.64}\right] = 4.04$$

Thus, the regression equation of X_3 on X_1 and X_2 is

$$X_3 - 74 = 4.36(X_1 - 6.8) + 4.04(X_2 - 7)$$

$\therefore$ $$X_3 = 16.07 + 4.36X_1 + 4.04X_2$$

Final marks of students who scored 9 and 7 marks :

When $X_1 = 9$ and $X_2 = 7$

$$X_3 = 16.07 + 4.36(9) + 4.04(7)$$
$$= 16.07 + 39.24 + 28.28 = 83.59 \text{ or } 84$$

Final marks of students who scored 4 and 8 marks

When $X_1 = 4$ and $X_2 = 8$

$$X_3 = 16.07 + 4.36(4) + 4.04(8)$$
$$= 16.07 + 17.44 + 32.32 = 65.8 \text{ or } 66$$

Example 6. **Given the following, determine the regression equations of :**

(i) x_1 **on** x_2 **and** x_3 **and**

(ii) x_2 **on** x_1 **and** x_3 **when the variates are measured from their means :**

$$r_{12} = 0 \cdot 8 \quad r_{13} = 0 \cdot 6 \quad r_{23} = 0 \cdot 5$$
$$\sigma_1 = 10, \sigma_2 = 8, \sigma_3 = 5$$

Solution. (i) The regression equation of x_1 on x_2 and x_3 when variates are measured from means is given by :

$$x_1 = b_{12.3}x_2 + b_{13.2}x_2 \quad \text{where, } x_1 = X_1 - \overline{X}_1, x_2 = X_2 - \overline{X}_2, x_3 = X_3 - \overline{X}_3$$

$$b_{12\cdot3}=\left[\frac{\sigma_1}{\sigma_2}\right]\times\left[\frac{r_{12}-r_{13}\cdot r_{23}}{1-r_{23}^2}\right]$$

$$=\left[\frac{10}{8}\right]\times\left[\frac{(0\cdot8)-(0\cdot6)(0\cdot5)}{1-(0\cdot5)^2}\right]=0\cdot833$$

$$b_{13\cdot2}=\left[\frac{\sigma_1}{\sigma_3}\right]\times\left[\frac{r_{13}-r_{12}\cdot r_{32}}{1-r_{32}^2}\right]$$

$$=\left[\frac{10}{5}\right]\times\left[\frac{(0\cdot6)-(0\cdot8)(0\cdot5)}{1-(0\cdot5)^2}\right]=0\cdot533$$

∴ The required regression equation is :

$$x_1=\cdot833\,x_2+\cdot533\,x_3$$

(ii) The regression equation of x_2 on x_1 and x_3 when variates are measured from means is given by :

$$x_2=b_{21\cdot3}\,x_1+b_{23\cdot1}\,x_3$$

$$b_{21\cdot3}=\left[\frac{\sigma_2}{\sigma_1}\right]\times\left[\frac{r_{21}-r_{23}\cdot r_{13}}{1-r_{13}^2}\right]$$

$$=\left[\frac{8}{10}\right]\times\left[\frac{(0\cdot8)-(\cdot5)(\cdot6)}{1-(\cdot6)^2}\right]=\cdot625$$

$$b_{23\cdot1}=\left[\frac{\sigma_2}{\sigma_3}\right]\times\left[\frac{r_{23}-r_{21}\cdot r_{31}}{1-r_{31}^2}\right]$$

$$=\left[\frac{8}{5}\right]\times\left[\frac{(\cdot5)-(0\cdot8)(0\cdot6)}{1-(\cdot6)^2}\right]=0\cdot05$$

∴ The required regression equation is :

$$x_2=\cdot625\,x_1+\cdot05\,x_3$$

Example 7. **A random sample of 15 students of Basic Statistics course when observed for weights (X_1), age (X_2) and height (X_3) offered the following information :**

$r_{12}=0\cdot8$, $r_{23}=0\cdot3$, $r_{13}=0\cdot5$, $S_1=8\cdot5$, $S_2=4\cdot5$, $S_3=2\cdot1$

$\overline{X}_1=70$ kg, $\overline{X}_2=22$ yrs and $\overline{X}_3=160$ cms.

Obtain :

(i) Multiple and partial correlation coefficients $R_{1\cdot23}$ and $r_{13\cdot2}$.

(ii) Multiple regression of X_1 on X_2 and X_3 and estimate the value of X_1 for $X_2=25$ yrs and $X_3=140$ cms.

Solution. (i) $$R_{1\cdot23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$

$$b_{12\cdot3}=\left[\frac{\sigma_1}{\sigma_2}\right]\times\left[\frac{r_{12}-r_{13}\cdot r_{23}}{1-r_{23}^2}\right]$$

$$=\left[\frac{10}{8}\right]\times\left[\frac{(0\cdot8)-(0\cdot6)(0\cdot5)}{1-(0\cdot5)^2}\right]=0\cdot833$$

$$b_{13\cdot2}=\left[\frac{\sigma_1}{\sigma_3}\right]\times\left[\frac{r_{13}-r_{12}\cdot r_{32}}{1-r_{32}^2}\right]$$

$$=\left[\frac{10}{5}\right]\times\left[\frac{(0\cdot6)-(0\cdot8)(0\cdot5)}{1-(0\cdot5)^2}\right]=0\cdot533$$

$\therefore$ The required regression equation is :

$$x_1=\cdot833x_2+\cdot533x_3$$

(ii) The regression equation of x_2 on x_1 and x_3 when variates are measured from means is given by :

$$x_2=b_{21\cdot3}x_1+b_{23\cdot1}x_3$$

$$b_{21\cdot3}=\left[\frac{\sigma_2}{\sigma_1}\right]\times\left[\frac{r_{21}-r_{23}\cdot r_{13}}{1-r_{13}^2}\right]$$

$$=\left[\frac{8}{10}\right]\times\left[\frac{(0\cdot8)-(\cdot5)(\cdot6)}{1-(\cdot6)^2}\right]=\cdot625$$

$$b_{23\cdot1}=\left[\frac{\sigma_2}{\sigma_3}\right]\times\left[\frac{r_{23}-r_{21}\cdot r_{31}}{1-r_{31}^2}\right]$$

$$=\left[\frac{8}{5}\right]\times\left[\frac{(\cdot5)-(0\cdot8)(0\cdot6)}{1-(\cdot6)^2}\right]=0\cdot05$$

$\therefore$ The required regression equation is :

$$x_2=\cdot625x_1+\cdot05x_3$$

Example 7. **A random sample of 15 students of Basic Statistics course when observed for weights (X_1), age (X_2) and height (X_3) offered the following information :**

$r_{12}=0\cdot8$, $r_{23}=0\cdot3$, $r_{13}=0\cdot5$, $S_1=8\cdot5$, $S_2=4\cdot5$, $S_3=2\cdot1$

$\overline{X}_1=70$ kg, $\overline{X}_2=22$ yrs and $\overline{X}_3=160$ cms.

Obtain :

(i) Multiple and partial correlation coefficients $R_{1\cdot23}$ and $r_{13\cdot2}$.

(ii) Multiple regression of X_1 on X_2 and X_3 and estimate the value of X_1 for $X_2=25$ yrs and $X_3=140$ cms.

Solution. (i) $$R_{1\cdot23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$

$$= \sqrt{\frac{(0\cdot8)^2 + (0\cdot5)^2 - 2(0\cdot8)(0\cdot5)(0\cdot3)}{1-(0\cdot3)^2}}$$

$$= \sqrt{\frac{0\cdot64 + 0\cdot25 - 0\cdot24}{0\cdot91}} = \sqrt{\frac{0\cdot65}{0\cdot91}} = 0\cdot8452$$

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,.\,r_{23}}{\sqrt{1-r_{13}^2}\,\sqrt{1-r_{23}^2}}$$

$$= \frac{(0\cdot8) - (0\cdot5)(0\cdot3)}{\sqrt{1-(0\cdot5)^2}\,\sqrt{1-(0\cdot3)^2}}$$

$$= \frac{0\cdot8 - 0\cdot15}{\sqrt{\cdot75}\,\sqrt{0\cdot91}} = \frac{0\cdot65}{0\cdot8261} = 0\cdot7868$$

(ii) Multiple Regression on X_1 on X_2 and X_3 is given by :

$$X_1 - \overline{X}_1 = b_{12\cdot3}(X_2 - \overline{X}_2) + b_{13\cdot2}(X_3 - \overline{X}_3)$$

$$b_{12\cdot3} = \left[\frac{S_1}{S_2}\right] \times \left[\frac{r_{12} - r_{13}\,.\,r_{23}}{1-r_{23}^2}\right]$$

$$= \left[\frac{8\cdot5}{4\cdot5}\right] \times \left[\frac{0\cdot8 - (0\cdot5)(0\cdot3)}{1-(\cdot3)^2}\right] = 1\cdot349$$

$$b_{13\cdot2} = \left[\frac{S_1}{S_3}\right] \times \left[\frac{r_{13} - r_{12}\,.\,r_{32}}{1-r_{32}^2}\right]$$

$$= \left[\frac{8\cdot5}{2\cdot1}\right] \left[\frac{(0\cdot5) - (0\cdot8)(0\cdot3)}{1-(0\cdot3)^2}\right] = 1\cdot156$$

Substituting the values in the equation, we get

$$X_1 - 70 = 1\cdot349(X_2 - 22) + 1\cdot156(X_3 - 160)$$

$$X_1 - 70 = 1\cdot349X_2 - 29\cdot678 + 1\cdot156X_3 - 184\cdot96$$

$$\therefore \quad X_1 = 1\cdot349X_2 + 1\cdot156X_3 - 144\cdot638$$

Estimation of X_1 for $X_2 = 25$ and $X_3 = 140$:

When $X_2 = 25$ and $X_3 = 140$, $X_1 = 1\cdot349(25) + 1\cdot156(140) - 144\cdot638$

$$= 33\cdot725 + 161\cdot84 - 144\cdot638 = 50\cdot927$$

Example 8. **In a trivariate distribution :**

$$\sigma_1 = 3,\ \sigma_2 = 4,\ \sigma_3 = 5$$

$$r_{23} = 0\cdot4,\ r_{31} = 0\cdot6,\ r_{12} = 0\cdot7$$

(i) Compute $r_{23\cdot1}$ and $R_{1\cdot23}$

(ii) Determine the regression equation of x_1 on x_2 and x_3 if the variates are measured from their means :

Solution. (i)

$$r_{23\cdot1} = \frac{r_{23} - r_{21}\,.\,r_{31}}{\sqrt{1-r_{21}^2}\,\sqrt{1-r_{31}^2}}$$

$$= \frac{(0\cdot4)-(0\cdot7)(0\cdot6)}{\sqrt{1-(0\cdot7)^2}\sqrt{1-(0\cdot6)^2}}$$

$$= \frac{0\cdot4-(\cdot42)}{\sqrt{0\cdot51}\sqrt{0\cdot64}} = \frac{-0\cdot02}{0\cdot5713} = -0\cdot035$$

$$R_{1\cdot23} = \sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$

$$= \sqrt{\frac{(0\cdot7)^2+(0\cdot6)^2-2(0\cdot7)(0\cdot6)(0\cdot4)}{1-(0\cdot4)^2}}$$

$$= \sqrt{\frac{0\cdot49+0\cdot36-0\cdot336}{0\cdot84}} = \sqrt{\frac{0\cdot514}{0\cdot84}} = 0\cdot782$$

(ii) The regression equation of x_1 on x_2 and x_3 when variates are measured from mean is given by :

$$x_1 = b_{12\cdot3}x_2 + b_{13\cdot2}x_3 \quad \text{where, } x_1 = X_1 - \overline{X}_1,\ x_2 = X_2 - \overline{X}_2,\ x_3 = X_3 - \overline{X}_3$$

$$b_{12\cdot3} = \left[\frac{\sigma_1}{\sigma_2}\right] \times \left[\frac{r_{12}-r_{13}\cdot r_{23}}{1-r_{23}^2}\right]$$

$$= \left[\frac{3}{4}\right]\left[\frac{(0\cdot7)-(0\cdot6)(0\cdot4)}{1-(0\cdot4)^2}\right]$$

$$= \frac{0\cdot75\times0\cdot46}{0\cdot84} = \frac{0\cdot345}{0\cdot84} = 0\cdot41$$

$$b_{13\cdot2} = \left[\frac{\sigma_1}{\sigma_3}\right] \times \left[\frac{r_{13}-r_{12}\cdot r_{32}}{1-r_{32}^2}\right]$$

$$= \left[\frac{3}{5}\right] \times \left[\frac{(0\cdot6)-(0\cdot7)(0\cdot4)}{1-(0\cdot4)^2}\right] = \frac{0\cdot6\times0\cdot32}{0\cdot84} = \frac{0\cdot192}{0\cdot84} = 0\cdot229$$

Thus, the required regression equation is :

$$x_1 = 0\cdot41x_2 + 0\cdot229x_3.$$

STANDARD ERROR OF ESTIMATE (OR RELIABILITY OF ESTIMATES) FOR MULTIPLE REGRESSION

The standard error of estimate measures the realiability of the estimates given by the multiple regression equation. It shows to what extent the estimated valued given by the regression equations are closer to the actual values.

For three regression equations, there are three standard error or estimates :

(1) Standard Error of Estimate of X_1 on X_2 and X_3 ($S_{1\cdot23}$)

(2) Standard Error of Estimate of X_2 on X_1 and X_3 ($S_{2\cdot13}$)

(3) Standard Error of Estimate of X_3 on X_1 and X_2 ($S_{3\cdot12}$)

The formulae for calculating the standard error of estimates are given as follows :

$$S_{1\cdot23} = \sigma_1 \cdot \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$S_{2\cdot13} = \sigma_2 \cdot \sqrt{\frac{1 - r_{21}^2 - r_{23}^2 - r_{13}^2 + 2r_{21} \cdot r_{23} \cdot r_{13}}{1 - r_{13}^2}}$$

$$S_{3\cdot12} = \sigma_3 \cdot \sqrt{\frac{1 - r_{31}^2 - r_{32}^2 - r_{13}^2 + 2r_{31} \cdot r_{32} \cdot r_{12}}{1 - r_{12}^2}}$$

Example 9. **If $r_{12} = 0{\cdot}8$, $r_{13} = 0{\cdot}5$, $r_{23} = 0{\cdot}3$ and $S_1 = 8{\cdot}5$, compute the standard error of estimate of X_1 on X_2 and X_3.**

Solution. Standard Error of Estimate of X_1 on X_2 and X_3 is given by :

$$S_{1\cdot23} = \sigma_1 \cdot \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= 8{\cdot}5 \sqrt{\frac{1 - (0{\cdot}8)^2 - (0{\cdot}5)^2 - (0{\cdot}3)^2 + 2(0{\cdot}8)(0{\cdot}5)(0{\cdot}3)}{1 - (0{\cdot}3)^2}} = 4{\cdot}543$$

Coefficient of Multiple Determination (R^2)

The coefficient of determination in multiple regression denoted by $R_{1\cdot23}^2$ is similar to the coefficient of determination r^2 in the simple linear regression. It represents the proportion (fraction) of the total variation in the dependent variable X_1 that has been explained by the independent variables (X_2 and X_3) in the multiple regression equation.

For example, if $R_{1\cdot23} = 0{\cdot}7252$, then $R_{1\cdot23}^2 = 0{\cdot}5259 = 0{\cdot}526$

The value of $R_{1\cdot23}^2 = 0{\cdot}526$ indicates that 52·6% variation in the dependent variable (X_1) are explained by the independent variables X_2 and X_3 in the multiple regression equation of X_1 on X_2 and X_3.

Example 10. **A random sample of 15 students of advanced course in statistics when observed for weight (X_1), age (X_2) and height (X_3) offered the following information :**

$r_{12} = 0{\cdot}8$, $r_{13} = 0{\cdot}5$, $r_{23} = 0{\cdot}3$

$S_1 = 8{\cdot}5$, $S_2 = 4{\cdot}5$ and $S_3 = 2{\cdot}1$

Find the following :

(a) Partial regression coefficient $b_{1\cdot23}$ and $b_{13\cdot2}$.

(b) Standard error of estimate $S_{1\cdot23}$.

(c) Correlation Coefficients $R_{1\cdot23}$ and $r_{12\cdot3}$.

(d) Multiple regression of X_1 on X_2 and X_3 when $\overline{X}_1 = 70$ kg, $\overline{X}_2 = 22$ years and $\overline{X}_3 = 150$ cm.

(e) Weight of a student (X_1) of 25 years of age and 140 cm in height.

Solution. Given : $r_{12} = 0{\cdot}8, r_{13} = 0{\cdot}5, r_{23} = 0{\cdot}3$

$$S_1 = 8{\cdot}5,\ S_2 = 4{\cdot}5,\ S_3 = 2{\cdot}1$$

(a) $$b_{12{\cdot}3} = \left[\frac{S_1}{S_2}\right] \cdot \left[\frac{r_{12} - r_{13} \cdot r_{23}}{1 - r_{23}^2}\right]$$

$$= \left[\frac{8{\cdot}5}{4{\cdot}5}\right] \times \left[\frac{0{\cdot}8 - (0{\cdot}5)\,(0{\cdot}3)}{1 - (0{\cdot}3)^2}\right] = 1{\cdot}349$$

$$b_{13{\cdot}2} = \left[\frac{S_1}{S_3}\right] \times \left[\frac{r_{13} - r_{12} \cdot r_{32}}{1 - r_{32}^2}\right]$$

$$= \left[\frac{8{\cdot}5}{2{\cdot}1}\right] \times \left[\frac{(0{\cdot}5) - (0{\cdot}8)\,(0{\cdot}3)}{1 - (0{\cdot}3)^2}\right] = 1{\cdot}156$$

(b) $$S_{1{\cdot}23} = S_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= 8{\cdot}5 \times \sqrt{\frac{1 - (0{\cdot}8)^2 - (0{\cdot}5)^2 - (0{\cdot}3)^2 + 2\,(0{\cdot}8)\,(0{\cdot}5)\,(0{\cdot}3)}{1 - (0{\cdot}3)^2}}$$

$$= 8{\cdot}5 \times \sqrt{\frac{1 - {\cdot}64 - {\cdot}25 - 0{\cdot}09 + 0{\cdot}24}{0{\cdot}91}} = 4{\cdot}543$$

(c) $$R_{1{\cdot}23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0{\cdot}8)^2 + (0{\cdot}5)^2 - 2\,(0{\cdot}8)\,(0{\cdot}5)\,(0{\cdot}3)}{1 - (0{\cdot}3)^2}}$$

$$= \sqrt{\frac{0{\cdot}64 + {\cdot}25 - 0{\cdot}24}{0{\cdot}91}} = \sqrt{\frac{0{\cdot}65}{0{\cdot}91}} = 0{\cdot}8452$$

$$r_{12{\cdot}3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

$$= \frac{(0{\cdot}8) - (0{\cdot}5) \cdot (0{\cdot}3)}{\sqrt{1 - (0{\cdot}5)^2}\ \sqrt{1 - (0{\cdot}3)^2}} = \frac{0{\cdot}65}{0{\cdot}8261} = 0{\cdot}7868$$

(d) Multiple Regression Equation of X_1 on X_2 and X_3 :

$$X_1 - \overline{X}_1 = b_{12{\cdot}3}\,(X_2 - \overline{X}_2) + b_{13{\cdot}2}\,(X_3 - \overline{X}_3)$$

Substituting the values, we have

$$X_1 - 70 = 1{\cdot}349\,(X_2 - 22) + 1{\cdot}156\,(X_3 - 150)$$

$$X_1 = -133{\cdot}078 + 1{\cdot}349\,X_2 + 1{\cdot}156\,X_3$$

(e) For $X_2 = 25$ and $X_3 = 140$,

$$X_1 = -133{\cdot}078 + 1{\cdot}348\,(25) + 1{\cdot}156\,(140)$$

$$= -133{\cdot}078 + 33{\cdot}725 + 161{\cdot}84 = 62{\cdot}487$$

Example 11. **Given the following data, determine the regression equation of x_1 on x_2 and x_3 if the variates are measured from their means :**

$$r_{12}=0\cdot8, \qquad r_{13}=0\cdot6, \qquad r_{23}=0\cdot5$$
$$\sigma_1=10, \qquad \sigma_2=8, \qquad \sigma_3=15$$

Also find the standard error of the estimate of x_1 on x_2 and x_3.

Solution. The regression equation of x_1 on x_2 and x_3 is :

$$x_1=b_{12\cdot3}\,x_2+b_{13\cdot2}\,x_3 \text{ where, } x_1=X_1-\bar{X}_1, x_2=X_2-\bar{X}_2 \text{ and } x_3=X_3-\bar{X}_3$$

Here,
$$b_{12\cdot3}=\left[\frac{\sigma_1}{\sigma_2}\right]\cdot\left[\frac{r_{12}-r_{13}\cdot r_{23}}{1-r_{23}^2}\right]$$
$$=\left[\frac{10}{8}\right]\times\left[\frac{0\cdot8-(0\cdot6)(0\cdot5)}{1-(0\cdot5)^2}\right]$$
$$=\left[\frac{10}{8}\right]\times\left[\frac{0\cdot8-\cdot30}{1-\cdot25}\right]=\frac{10}{8}\times\frac{0\cdot50}{0\cdot75}=0\cdot833$$
$$b_{13\cdot2}=\frac{\sigma_1}{\sigma_3}\cdot\left[\frac{r_{13}-r_{12}\cdot r_{32}}{1-r_{32}^2}\right]$$
$$=\frac{10}{5}\times\left[\frac{(0\cdot6)-(0\cdot8)(0\cdot5)}{1-(0\cdot5)^2}\right]$$
$$=\frac{10}{5}\times\frac{0\cdot20}{0\cdot75}=\frac{2}{3\cdot75}=0\cdot53$$

Thus, regression equation of x_1 on x_2 and x_3 is :

$$x_1=0\cdot833\,x_2+0\cdot53\,x_3$$

Standard Error of Estimate of X_1 on X_2 and X_3

$$S_{1\cdot23}=\sigma_1\sqrt{\frac{1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$
$$=10\cdot\sqrt{\frac{1-(0\cdot8)^2-(0\cdot6)^2-(0\cdot5)^2-2(0\cdot8)(0\cdot6)(0\cdot5)}{1-(0\cdot5)^2}}$$
$$=10\cdot\sqrt{\frac{1-\cdot64-\cdot36-\cdot25+\cdot48}{0\cdot75}}$$
$$=10\cdot\sqrt{\frac{0\cdot23}{0\cdot75}}=10\times\cdot5537=5\cdot537$$

Example 11A. **The following values have been obtained from the measurement of three variables x_1, x_2 and x_3 :**

$$\bar{X}_1=6\cdot8 \qquad \bar{X}_2=7\cdot0 \qquad \bar{X}_3=7\cdot4$$
$$S_1=1\cdot0 \qquad S_2=0\cdot80 \qquad S_3=0\cdot90$$
$$r_{12}=0\cdot60 \qquad r_{13}=0\cdot70 \qquad r_{23}=0\cdot65$$

(i) Obtain regression eqaution of X_1 on X_2 and X_3.
(ii) Estimate the value of X_1 for $X_2 = 10$ and $X_3 = 9$.
(iii) Find the coefficient of multiple determination $R^2_{1\cdot23}$ from r_{12} and $r_{13\cdot2}$.

Solution. The regression equation of X_1 on X_2 and X_3 is given by :

$$X_1 - \overline{X}_1 = b_{12\cdot3}(X_2 - \overline{X}_2) + b_{13\cdot2}(X_3 - \overline{X}_3) \quad ...(i)$$

where,

$$b_{12\cdot3} = \frac{s_1}{s_2}\left[\frac{r_{12} - r_{13}\cdot r_{23}}{1 - r^2_{23}}\right] = \frac{1}{0\cdot80}\left[\frac{0\cdot60 - 0\cdot70 \times 0\cdot65}{1 - (0\cdot65)^2}\right]$$

or $$b_{12\cdot3} = (1\cdot25)\left[\frac{0\cdot60 - 0\cdot455}{0\cdot578}\right] = 0\cdot313$$

$$b_{13\cdot2} = \frac{s_1}{s_2}\cdot\left[\frac{r_{13} - r_{12}\cdot r_{32}}{1 - r^2_{32}}\right] = \frac{1}{0\cdot90}\left[\frac{0\cdot70 - 0\cdot60 \times 0\cdot65}{1 - (0\cdot65)^2}\right]$$

$$= (1\cdot111)\left[\frac{0\cdot70 - 0\cdot39}{0\cdot578}\right] = 0\cdot595$$

Substituting the values in equation (i), we have,

$$X_1 - 6\cdot8 = 0\cdot313(X_2 - 7\cdot0) + 0\cdot595(X_3 - 7\cdot4)$$

or $$X_1 = 0\cdot206 + 0\cdot313X_2 + 0\cdot595X_3$$

(ii) Substituting for $X_2 = 10$ and $X_3 = 9$ in the above regression and solving for x_1.

$$X_1 = 0\cdot206 + 0\cdot313(10) + 0\cdot595(9) = 8\cdot691$$

(iii) Multiple and partial correlation coefficients are related as :

$$R^2_{12\cdot3} = 1 - (1 - r^2_{12})(1 - r^2_{13\cdot2})$$

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\cdot r_{32}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{32}}} = \frac{0\cdot70 - 0\cdot60 \times 0\cdot65}{\sqrt{1 - (0\cdot60)^2}\sqrt{1 - (0\cdot65)^2}}$$

$$= \frac{0\cdot70 - 0\cdot39}{0\cdot8 \times 0\cdot760} = 0\cdot509$$

or, $$r^2_{13\cdot2} = 0\cdot259$$

Substituting the values of r^2_{12} and $r^2_{13\cdot2}$ for $R^2_{1\cdot23}$ we have

$$R^2_{1\cdot23} = 1 - (1 - 0\cdot36)(1 - 0\cdot259) = 0\cdot526$$

EXERCISE – 4

1. The following constants are obtained from measurements of length in mm (x_1), volume in c.c. (x_2) and weight in gm (x_3) of 300 eggs :

$\overline{X}_1 = 55\cdot95$ $S_1 = 2\cdot26$ $r_{12} = 0\cdot578$
$\overline{X}_2 = 51\cdot48$ $S_2 = 4\cdot39$ $r_{13} = 0\cdot581$
$\overline{X}_3 = 56\cdot03$ $S_3 = 4\cdot41$ $r_{23} = 0\cdot974$

Obtain the linear
estimate the wei

2. In a trivarlate di
$\sigma_1 = 2\cdot7$
$r_{12} = 0\cdot2$
Determine the r
their means.
3. Given the follow
$\overline{X}_1 = 6,$
$\sigma_1 = 1,$
$r_{12} = 0\cdot6$
Obtain the linea
and $X_2 = 5$.
4. The following r

Find (i) partial c
the partial regre

5. If $r_{12} = 0\cdot926, r_1$
of X_1 on X_2 and
6. A random samp
height (x_3) offe
$r_{12} = 0\cdot$
$S_1 = 5\cdot6$
Obtain the follo
(a) Partial regr
(b) Standard e
(c) Coefficient
(d) Coefficient

7. Given the follo
variates measu
$r_{12} = 0\cdot$
$\sigma_1 = 10$
Also find the st

8. Given the follo
$\overline{X}_1 = 30$
$\overline{X}_2 = 35$
$\overline{X}_3 = 40$
Also calculate

Obtain the linear regression equation of egg weight on egg length and egg volume. Hence estimate the weight of an egg whose length is 58 mm and volume is 52·5 c.c.

[Ans. $X_3 = 3{\cdot}54 + 0{\cdot}052X_1 + 0{\cdot}963X_2$, $X_3 = 57{\cdot}11$ gms.]

2. In a trivarlate distribution :

$\sigma_1 = 2{\cdot}7$, $\sigma_2 = 2{\cdot}4$, $\sigma_3 = 2{\cdot}7$

$r_{12} = 0{\cdot}28$, $r_{23} = 0{\cdot}49$, $r_{31} = 0{\cdot}51$

Determine the regression equation of x_3 on x_1 and x_2 if the variates are measured from their means.

[Ans. $x_3 = 0{\cdot}405x_1 + 0{\cdot}424x_2$]

3. Given the following data :

$\overline{X}_1 = 6$, $\overline{X}_2 = 7$, $\overline{X}_3 = 8$

$\sigma_1 = 1$, $\sigma_2 = 2$, $\sigma_3 = 3$

$r_{12} = 0{\cdot}6$, $r_{13} = 0{\cdot}7$, $r_{23} = 0{\cdot}8$

Obtain the linear regression equation of X_3 on X_1 and X_3. Hence estimate X_3 when $X_1 = 4$ and $X_2 = 5$.

[Ans. $x_3 = -4{\cdot}41 + 1{\cdot}03x_1 + 0{\cdot}89x_2$, $x_3 = 4{\cdot}16$]

4. The following results were obtained in the analysis of a trivariate distribution ;

$S_1 = 3$, $S_2 = S_3 = 5$, $r_{12} = 0{\cdot}7$, $r_{23} = r_{31} = 0{\cdot}6$

Find (i) partial correlation coefficient $r_{12{\cdot}3}$ (ii) Multiple correlation coefficient $R_{1{\cdot}23}$ and (iii) the partial regression coefficients ($b_{12{\cdot}3}$ and $b_{13{\cdot}2}$).

[Ans. $r_{12{\cdot}3} = 0{\cdot}531$, $R_{1{\cdot}23} = 0{\cdot}735$, $b_{12{\cdot}3} = 0{\cdot}319$, $b_{13{\cdot}2} = 0{\cdot}169$]

5. If $r_{12} = 0{\cdot}926$, $r_{13} = 0{\cdot}891$, $r_{23} = 0{\cdot}955$ and $S_1 = 1{\cdot}51$, compute the standard error of estiamte of X_1 on X_2 and X_3 ($S_{1{\cdot}23}$).

[Ans. $S_{1{\cdot}23} = 0{\cdot}5702$]

6. A random sample of 50 students of M. Com. when observed for weight (x_1), age (x_2), and height (x_3) offered the following information :

$r_{12} = 0{\cdot}7$, $r_{13} = 0{\cdot}8$, $r_{23} = 0{\cdot}5$

$S_1 = 5{\cdot}6$, $S_2 = 4{\cdot}5$, $S_3 = 3{\cdot}5$

Obtain the following :

(a) Partial regression coefficient $b_{12{\cdot}3}$ and $b_{13{\cdot}2}$.

(b) Standard error of estimate $S_{1{\cdot}23}$.

(c) Coefficient of multiple correlation ($R_{1{\cdot}23}$).

(d) Coefficient of partial correlation ($r_{12{\cdot}3}$).

[Ans. $b_{12{\cdot}3} = 0{\cdot}496$, $b_{13{\cdot}2} = 0{\cdot}96$, $S_{1{\cdot}23} = 2{\cdot}74$, $R_{1{\cdot}23} = 0{\cdot}87$, $r_{12{\cdot}3} = 0{\cdot}516$]

7. Given the following data, determine the regression equation of x_1 on x_2 and x_3 if the variates measured from their means :

$r_{12} = 0{\cdot}8$, $r_{13} = 0{\cdot}6$, $r_{23} = 0{\cdot}5$

$\sigma_1 = 10$, $\sigma_2 = 8$, $\sigma_3 = 5$

Also find the standard error of estimate of x_1 on x_2 and x_3.

[Ans. $x_1 = 0{\cdot}833x_2 + 0{\cdot}53x_2$, $S_{1{\cdot}23} = 5{\cdot}537$]

8. Given the following data, calculate the estimated value of X_1 when $X_2 = 20$ and $X_3 = 25$.

$\overline{X}_1 = 30$ $S_1 = 5$ $r_{12} = -0{\cdot}4$

$\overline{X}_2 = 35$ $S_2 = 10$ $r_{13} = -0{\cdot}5$

$\overline{X}_3 = 40$ $S_3 = 15$ $r_{23} = 0{\cdot}6$

Also calculate the standard error of estimate of X_1 on X_2 and X_3.

[Ans. $X_3 = 38{\cdot}225 - 0{\cdot}075X_2 - 0{\cdot}14X_3$, $X_1 = 33$, $S_{1{\cdot}23} = 1{\cdot}1$]

9. Given the following data :

$\bar{X}_1 = 55$ $S_1 = 5$ $r_{12} = 0{\cdot}57$

$\bar{X}_2 = 51$ $S_2 = 7$ $r_{13} = 0{\cdot}58$

$\bar{X}_3 = 56$ $S_3 = 9$ $r_{23} = 0{\cdot}97$

Calculate :

(i) Multiple Regression of X_3 on X_1 and X_2.

(ii) Multiple Correlation coefficient $R_{1{\cdot}23}$

[**Ans.** $X_3 = 0{\cdot}08X_1 + 1{\cdot}21X_2 - 10{\cdot}11$, $R_{1{\cdot}23} = 0{\cdot}97$]

10. From the following data,

$r_{12} = 0{\cdot}8, r_{13} = 0{\cdot}5, r_{23} = 0{\cdot}3$

$S_1 = 8{\cdot}5, S_2 = 4{\cdot}5$ and $S_3 = 2{\cdot}1$

(i) Obtain the regression equation of X_1 on X_2 and X_3 with $\bar{X}_1 = 70$ kg, $\bar{X}_2 = 22$ years, and $\bar{X}_3 = 150$ cm.

(ii) Estimate the value of X_1 on $X_2 = 25$ years and $X_3 = 140$ cm, and

(iii) Find the coefficient of multiple determination $R^2_{1{\cdot}23}$ from r_{12} and $r_{13{\cdot}2}$. What does R^2 indicate ?

[**Ans.** $X_1 = 1{\cdot}349X_2 + 1{\cdot}156X_3 - 133{\cdot}078$, $X_1 = 62{\cdot}487$ $R^2_{1{\cdot}23} = 0{\cdot}7443$, R^2 indicates the 74·43% variation in X_1 are explained by the multiple regression equation]

MISCELLANEOUS SOLVED EXAMPLE

Example 12. **In a trivariate distribution :**

$\bar{X}_1 = 28{\cdot}02$, $\bar{X}_2 = 4{\cdot}91$, $\bar{X}_3 = 594$, $S_1 = 4{\cdot}4$, $S_2 = 1{\cdot}1$, $S_3 = 80$

$r_{12} = 0{\cdot}80$, $r_{23} = -0{\cdot}56$, $r_{31} = -0{\cdot}40$

(i) Find the correlation coefficient $r_{23{\cdot}1}$ and $R_{1{\cdot}23}$.

(ii) Estimate the value of X_1 when $X_2 = 6{\cdot}0$ and $X_3 = 650$.

Solution.

(i) $$r_{23{\cdot}1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

Substituting the values, we get

$$r_{23{\cdot}1} = \frac{-0{\cdot}56 - (0{\cdot}80) \cdot (-0{\cdot}40)}{\sqrt{1 - (0{\cdot}80)^2}\sqrt{1 - (-0{\cdot}40)^2}}$$

$$= \frac{-0{\cdot}56 + {\cdot}32}{\sqrt{1 - {\cdot}64}\sqrt{1 - {\cdot}16}} = \frac{-0{\cdot}24}{0{\cdot}6 \times 0{\cdot}916} = -0{\cdot}436$$

$$R_{1{\cdot}23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 + 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

Substituting values, we get

$$R_{1\cdot23}=\sqrt{\frac{(0\cdot80)^2+(-0\cdot40)^2-2(0\cdot80)(-0\cdot40)(-0\cdot56)}{1-(-\cdot56)^2}}$$

$$=\sqrt{\frac{0\cdot64+\cdot16-\cdot3584}{1-\cdot3136}}=\sqrt{\frac{0\cdot4416}{0\cdot6864}}=0\cdot802$$

(ii) The linear regression equation of X_1 on X_2 and X_3 is :

$$X_1-\bar{X}_1=b_{12\cdot3}(X_2-\bar{X}_2)+b_{13\cdot2}(X_3-\bar{X}_3)$$

$$b_{12\cdot3}=\frac{S_1}{S_2}\times\left[\frac{r_{12}-r_{13}\cdot r_{23}}{1-r_{23}^2}\right]$$

$$=\frac{4\cdot4}{1\cdot1}\cdot\left[\frac{0\cdot80-0\cdot224}{0\cdot6864}\right]=\frac{4\cdot4}{1\cdot1}\cdot\left[\frac{0\cdot576}{0\cdot6864}\right]=3\cdot357$$

$$b_{13\cdot2}=\frac{S_1}{S_3}\cdot\left[\frac{r_{13}-r_{12}\cdot r_{32}}{1-r_{32}^2}\right]=\frac{4\cdot4}{80}\cdot\left[\frac{-0\cdot40-(0\cdot80)(-0\cdot56)}{1-(-\cdot56)^2}\right]$$

$$=\frac{4\cdot4}{80}\cdot\left[\frac{-0\cdot40+0\cdot448}{0\cdot6884}\right]=\frac{4\cdot4}{80}\cdot\left[\frac{0\cdot048}{0\cdot6864}\right]=0\cdot0038=0\cdot004$$

Substituting the values in the equation, we get

$$X_1-28\cdot02=3\cdot357(X_2-4\cdot91)+\cdot004(X_3-594)$$

or $$X_1-28\cdot02=3\cdot357X_2-16\cdot4828+\cdot004X_3-2\cdot376$$

$$X_1=3\cdot357X_2+\cdot004X_3+9\cdot1612$$

(iv) Estimation of X_1 for X_2 and X_3 :

When $X_2=6\cdot00, X_3=650$, $X_1=3\cdot357(6\cdot00)+\cdot004(650)+9\cdot1612$

$=20\cdot142+2\cdot6+9\cdot1612$

$=31\cdot9032$

IMPORTANT FORMULAE

Multiple Correlation Coefficients :

$$R_{1\cdot23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$

$$R_{2\cdot13}=\sqrt{\frac{r_{21}^2+r_{23}^2-2r_{21}\cdot r_{23}\cdot r_{13}}{1-r_{13}^2}}$$

$$R_{3\cdot12}=\sqrt{\frac{r_{31}^2+r_{32}^2-2r_{31}\cdot r_{32}\cdot r_{12}}{1-r_{12}^2}}$$

Partial Correlation Coefficients :

$$r_{12\cdot3}=\frac{r_{12}-r_{13}\cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$r_{23\cdot1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

$$r_{13\cdot2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$$

Relationship between Simple, Partial and Multiple Correlation Coefficients

$$1 - R_{1\cdot23}^2 = (1-r_{12}^2)(1-r_{13\cdot2}^2)$$

$$1 - R_{2\cdot13}^2 = (1-r_{21}^2)(1-r_{23\cdot1}^2) \quad \text{and}$$

$$1 - R_{3\cdot12}^2 = (1-r_{21}^2)(1-r_{32\cdot1}^2)$$

Multiple Regression of X_1 on X_2 and X_3 :

$$X_1 = a_{1\cdot23} + b_{12\cdot3} X_2 + b_{13\cdot2} X_3$$

Where, $$b_{12\cdot3} = \left[\frac{\sigma_1}{\sigma_2}\right] \times \left[\frac{r_{12} - r_{13} \cdot r_{23}}{1-r_{23}^2}\right]$$

Standard Error of Estimate :

$$S_{1\cdot23} = \sigma_1 \cdot \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12} \cdot r_{13} \cdot r_{23}}{1-r_{23}^2}}$$

QUESTIONS

1. Distinguish between partial correlation and multiple correlation.
2. Write down the expression for $r_{12\cdot3}$ and $R_{1\cdot23}$ in terms of r_{12}, r_{13} and r_{23}. Also state the limits within which $r_{12\cdot3}$ and $R_{1\cdot23}$ must lie.
3. Explain the concept of multiple regression and discuss its utility in business.
4. How will you fit a multiple regression equation of X_1 on X_2 and X_3 ?
5. Write the normal equations in case of multiple linear regression of X_1 on X_2 and X_3.
6. Write a short note on "Standard Error of Estiamte" for multiple regression.
7. Define simple, partial and multiple correlation coefficients and find their relationship with one another.
8. Write the equations/formulae to calculate the followings :
   (i) Regression equation for X_2 on X_1 and X_3.
   (ii) Partial regression coefficients ($b_{12\cdot3}$ and $b_{13\cdot2}$).
   (iii) Standard error of estimate of X_1 on X_2 and X_3 ($S_{1\cdot23}$)
   (iv) Multiple correlation coefficient ($R_{1\cdot23}$)
   (v) Partial correlation coefficient ($r_{12\cdot3}$)
9. How will you interpret the value of R^2 in a multiple regression equation ?

❀❀❀

2

Sampling & Sampling Distribution

INTRODUCTION

In all the spheres of life (such as Economic, Social and Business) the need for statistical investigation and data analysis is rising day by day. There are two methods of collection of statistical data : (i) **Census Method,** and (ii) **Sample Method.** Under census method, information relating to the entire field of investigation or units of population is collected; whereas under sample method, rather than collecting information about all the units of population, information relating to only selected units is collected. Before we make a detailed study of both the methods, we will explain some basic concepts related to them.

SOME BASIC CONCEPTS

(1) Universe or Population : In statistics, universe or population means an aggregate of items about which we obtain information. A universe or population means the entire field under investigation about which knowledge is sought. For example, if we want to collect information about the average monthly expenditure of all the 2,000 students of a college, then the entire aggregate of 2,000 students will be termed as Universe or Population. A population can be of two kinds (i) Finite and (ii) Infinite. In a finite population, number of items is definite such as, number of students or teachers in a college. On the other hand, an infinite population has infinite number of items *e.g.*, number of stars in the sky, number of water drops in an ocean, number of leaves on a tree or number of hairs on the head.

(2) Sample : A part of population is called sample. In other words, selected or sorted units from the population is known as a sample. In fact, a sample is that part of the population which we select for the purpose of investigation. For example, if an investigator selects 200 students from 2000 students of a college who represent all of them, then these 200 students will be termed as a sample. Thus, sample means some units selected out of a population which represent it.

CENSUS AND SAMPLE METHODS

There are two methods to collect statistical data :

(1) Census Method

(2) Sample Method

(1) Census Method

Census method is that method in which information or data is collected from each and every unit of the population relating to the problem under investigation and conclusions are drawn on their basis. This method is also called as **Complete Enumeration Method.** For example, suppose some information (like Monthly Expenditure, Average Height, Average Weight, etc.) is to be

collected regarding 2000 students of a college. For that purpose if we collect data by inquiring each and every student of the college then this method will be called as Census method. In this example, the whole college *i.e.*, all 2000 students will be considered as a population and every student as an individual will be called the unit of the population. Population in India is conducted after every ten years by using census method.

Merits and Demerits of Census Method

Merits

(i) **Reliable and Accurate Data :** Data obtained by census method have more reliability and accuracy because in this method data are collected by contacting each and every unit of the universe.

(ii) **Extensive Information :** This method gives detailed information about each unit of the universe. For example, Indian population census does not only provide the knowledge about the number of persons but also information about their age, occupation, income, education, marital status, etc.

(iii) **Suitability :** This method is more suitable for the population with limited scope and diverse characteristics. Use of this method is also appropriate where intensive study is desired.

Demerits

(i) **More Expensive :** Census method is an expensive one. More money is needed for it as information is collected from each unit of the population. This is why this method is used by Government mostly for very important issues like Census, etc.

(ii) **More Time :** This method involves much time for data collection because data are collected from each and every unit of the population. This results in delay in making statistical inferences.

(iii) **More Labour :** This method of data collection also involves very much labour. For this the enumerators in a large number are required.

(iv) **Not Suitable for Specific Problems :** This method is not suitable relating to certain specific problems and infinite population. For example, if the population is infinite or items of the population are perishable or very complex type, then the census method is not suitable.

(2) Sampling Method

Sampling method is that method in which data is collected from the sample of items selected from population and conclusions are drawn from them. For example, if a study is to be made regarding the monthly expenditure of 2000 students of a college, then instead of collecting information from each student of the college, if we collect information by selecting some students like 100, then this will be called Sampling Method. On the basis of sampling method, it is possible to study the monthly expenditure of all the students of the college. Sampling method has three main stages (i) to select a sample (ii) to collect information from it and (ii) to make inferences regarding the population.

Importance of Sampling Method

In modern times sampling method is an important and popular method of statistical inquiry. Besides economic and business world, this method is widely used in daily life. For example, a housewife comes to know of the coating of the whole lot of rice by observing two-three grains only. A doctor tests the blood of a patient by examining one or two drops of blood only. In the same way, we learn about the quality of a commodity while buying the items of daily use like wheat, rice, pulses, etc. by observing the sample or specimen. In factories, statistical quality controller inspects the quality of items by examining a few units produced. A teacher gets the knowledge about the

efficacy of his teaching by putting questions to a few students. In reality, there is no unit table of 10,000 left where sampling method is not used. -II, all the items 's tables, in a sample

Merits and Demerits of Sampling Method

Merits

(i) **Saving of Time and Money :** Sampling method is less expensive. It saves money and labour because only a few units of the population are studied.

(ii) **Saving of Time :** In sampling method, data can be collected more quickly as these are obtained from some items of the universe. Thus much time is saved.

(iii) **Intensive Study :** As number of items is less in sampling method, they can be intensively studied.

(iv) **Organisational Convenience :** In this method, research work can be organised and executed more conveniently. More skilled and competent investigators can be appointed.

(v) **More Reliable Results :** If sample is selected in such a manner as it represents totally the universe, then the results derived from it will be more accurate and reliable.

(vi) **More Scientific :** Sampling method is more scientific because data can be inquired with other samples.

(vii) **Only Method :** In some fields where inquiry by census method is impossible, then in such situation, sampling method alone is more appropriate. If the population is infinite or too widespread or of perishable nature, then sampling method is used in such cases.

Demerits

(i) **Less Accurate :** Sampling method has less accuracy because rather than making inquiry about each unit of the universe, partial inquiry or inquiry relating to some selected units only is made.

(ii) **Wrong Conclusions :** If method of selecting a sample is not unbiased or proper caution has not been taken, then results are definitely misleading.

(iii) **Less Reliable :** Compared to census method, there is more likelihood of the bias of the investigator, which makes the results less reliable.

(iv) **Need of Specified Knowledge :** This is a complex method as specialised knowledge is required to select a sample.

(v) **Not Suitable :** If all units of a population are different from one another, then sampling method will not prove to be much useful.

Difference between Census and Sample Method

The main difference between the census method and the sampling method are as follows :

(i) **Scope :** In census method, all items relating to a universe are investigated whereas in sampling method only a few items are inquired.

(ii) **Cost :** Census method is expensive from the point of view of time, money and labour whereas Sampling method economises on them.

(iii) **Field of Investigation :** Census method isused in investigations with limited field whereas sampling method is used for investigations with large field.

(iv) **Homogeneity :** Census method is useful where units of the population are heterogeneous whereas sampling method proves more useful where population units are homogenous.

(v) **Type of Universe :** In such fields where study of each and every unit of the universe is necessary, census method is more appropriate. On the contrary, when population is infinite or vast or liable to be destroyed as a result of complete enumeration, then sampling method is considered to be more appropriate.

SAMPLING METHODS

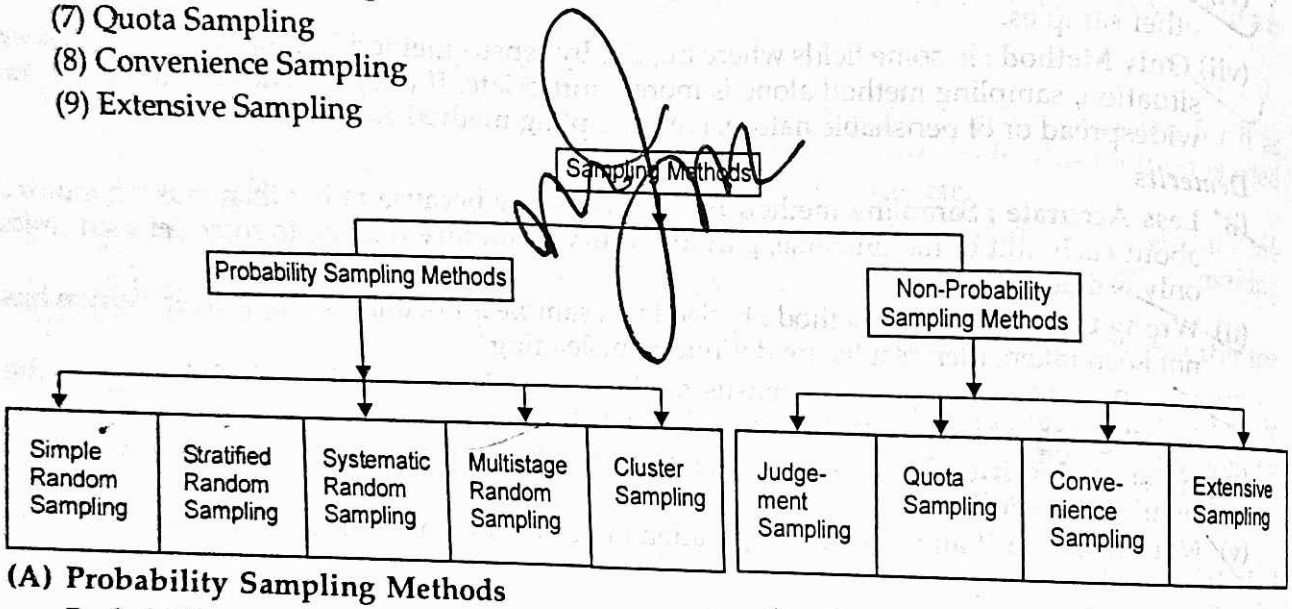
The method of selecting a sample out of a given population is called sampling. In other wodrs, sampling denotes the selection of a part of the aggregate statistical material with a view of obtaining information about the whole. Now a days, there are various methods of selecting sample from a population in accordance with various needs.

(A) Probability Sampling Methods :

(1) Simple Random Sampling
(2) Stratified Random Sampling
(3) Systematic Random Sampling
(4) Multistage Random Sampling
(5) Cluster Sampling

(B) Non-Probability Sampling Methods :

(6) Judgement Sampling
(7) Quota Sampling
(8) Convenience Sampling
(9) Extensive Sampling



(A) Probability Sampling Methods

Probability sampling methods are such methods of selecting a sample from the population in which all units of the universe are given equal chances of being included in the sample.

There are various variants of probability sampling methods, which are given below :

(1) **Simple Random Sampling :** Simple random sampling is that method in which each item of the universe has an equal chance of being selected in the sample. Which item will included in the sample and which not, such decision is not made by an investigator on his will but selection of the units is left on chance. According to random sampling, there are two methods of selecting a random sample:

(i) **Lottery Method :** In this method, each unit of the population is named or numbered which is marked on separate piece of paper. Such chits are folded and put into some urn or bag. Thereafter as many chits are made selected by some person as many units are to be included in a sample.

(ii) **Tables of Random Numbers :** Some experts has constructed random number tables. These tables help in selection of a sample. Of all such various tables, Tippett's Tables

are most famous and are in use. Tippett has constructed a four-digit table of 10,400 numbers by using numbers as many as 41,600. In this method, first of all, all the items of a population are written serially. Thereafter by making use of Tippett's tables, in accordance with the size of the sample, numbers are selected. The selection of a sample with the help of Tippett's table can be made clear by an example :

An Extract of Tippett's Table

2952	6641	3992	9792	7979	5911
3170	5524	4167	9525	1545	1396
7203	4356	1300	2693	2370	7483
3408	2762	3563	6107	6913	7691
0560	5246	1112	9025	6008	8127

For example, suppose 12 units are to be chosen out of 5000 units. With Tippett's table, to decide such units, firstly 5000 units will be serially ordered from 1 to 5000 and then from Tippett's table, 12 numbers will be chosen from the beginning which are less than 5000. These 12 numbers are follows :

2952	4167	4356	2370
3992	1545	1300	3408
3170	1396	2693	2762

The items of such serial numbers will be included in the sample. If units of the population are less than 100, then 4 digit random numbers will be made compact into two digit numbers, and then such two digit numbers will be selected. Like as to select 6 units out of 60 units, the units with serial numbers 29, 39, 31, 41, 15 and 13 will be included in the sample.

Merits

(i) There is no possibility of personal prejudice in this method. In other words, this method is free from personal bias.

(ii) Under this method, every unit of the universe gets the equal chance of being selected.

(iii) The use of this method saves time, money and labour.

Demerits

(i) If sample size is small, then sample is not adequately represented.

(ii) If universe is very small, then this method is not suitable.

(iii) If some items of the universe are so important that their inclusion in the sample is very essential, then this method will not be appropriate.

(iv) This method will not be appropriate when population has units with diverse characteristics.

(2) Stratified Random Sampling : This method is used when units of the universe are heterogeneous rather than homogeneous. Under this method, first of all units of the population are divided into different *strata* in accordance with their characteristics. Thereafter by using random sampling, sample items are selected from each stratum. For example, if 150 students are to be selected out of 1500 students of a college, then firstly the college students will be divided into three groups on the basis of Arts, Commerce and Science. Suppose there are 500, 700, 300 students respectively in three faculties and 10% sample is to be taken, then on the basis of random sampling 50, 70 and 30 students respectively will be selected by using random sampling. Thus, this method assumes equal representation to each class or group and all the units of the universe get equal chance of being selected in the sample.

Merits

(i) There is more likelihood of representation of units in this method.

(ii) Comparative study on the basis of facts at different strata is possible under this method.

(iii) This method has more accuracy.

Demerits

(i) This method has limited scope because this method can be adopted only when the population and its different strata are known.

(ii) There can be the possibility of prejudice if the population is not properly stratified.

(iii) If the population is too small in size, it is difficult to stratify it.

(3) Systematic Random Sampling: In this method, all the items of the universe are systematically arranged and numbered and then sample units are selected at equal intervals. For example, if 5 out of 50 students are to be selected for a sample, then 50 students would be numbered and systematically arranged. One item of the first 10 would be selected at random. Subsequently, every 10th item from the selected number will be selected to frame a samplé. If the first selected number is 5th item, then the subsequent numbers would be 15th, 25th, 35th and 45th.

Merits

(i) It is a simple method. Samples can be easily obtained by it.

(ii) This method involves very little time in sample selection and results are almost accurate.

Demerits

(i) In this method, each unit does not stand the equal chances of being selected because only the first unit is selected on random sampling basis.

(ii) If all the units are different in characteristics, then results will not be appropriate.

(4) Multistage Random Sampling : When sampling procedure passes through many stages, then it is known as multi-stage sampling. In this method, firstly the entire universe or population is divided into stages or substages. From the each stage some units are selected on random sampling basis. Thereafter these units are subdivided and on the basis of random sampling again some sub-units are selected. Thus, this goes on with sub-division further and selection on. For example, for the purpose of a study regarding Adult Education in Haryana State, first some districts will be selected on random basis. Thereafter out of the selected districts, some tehsils and out of tehsils, some villages or towns may be thus selected, further out of the villages or towns, some neighbourhood, or wards and out of the wards, some households will be selected from whom the inquiry will be made concerning the problem at hand.

Merits

(i) This is the best method of studying a universe or population on regional basis.

(ii) This method is suitable for those problems where decisions on the basis of sample alone can not be taken.

Demerits

(i) This method requires many tests to correctly estimate the level of accuracy which involves a lot of time and labour.

(ii) In this method, level of estimated accuracy level is predecided which does not seem logical.

(5) Cluster Sampling : In this method, simply the universe is divided into many groups called cluster and out of which a few clusters are selected on random basis and then the clusters are complete enumerated. This method is usually applied in industries like as in pharmaceutical

industry, a machine produces medicines tablets in the batches of hundred each, then for quality inspection, a few randomly selected batches are examined.

(B) Non-Probability Sampling Methods

Non-probability sampling methods are those methods in which selection of units is made on the basis of convenience or judgement of the investigator rather than on the basis of probability or chance. In such methods, selection of units is made in accordance with the specific objectives and convenience of the investigator.

(6) Judgement Sampling : Under this method, the selection of the sample items depends exclusively on the judgement of the investigator. In other words, the investigator exercises his judgement in the choice and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics under study. For example, if a sample of 20 students is to be selected from a class of 80 students for analysing the spending habits of the 10 students, the investigator would select 20 students, who in his opinion are representative of the class.

Merits

(i) This method is less expensive.

(ii) This method is very simple and easy.

(iii) This method is useful in those fields where almost similar units exist or some units are too important to be left out of the sample.

Demerits

(i) There is greater chance of the investigator's own prejudice in this method.

(ii) This method is not very accurate and reliable.

(7) Quota Sampling : In this method, the investigators are assigned definite quotas according to some criteria. They are instructed to obtain the required number to fill in each quota. The investigators select the individuals (*i.e.*, sample items) to collect information on their personal judgements within the quotas. When all or a part of the whole quota is not available or approachable, the quota is completed by supplementing new responds. Quota sampling is a type of judgement sampling.

Merits

(i) In this method, there is greater chance of important units being included.

(ii) Statistical inquiry is more organised in this method on account of the units of the quotas being fixed.

Demerits

(i) Possibility of prejudice shall remain.

(ii) There is greater likelihood of sampling error in this method.

(8) Convenience Sampling : In this type of non-probability sampling, the choice of the sample is left completely to the convenience of the investigator. The investigator obtain a sample according to his convenience. For example, a book publisher selects some teachers conveniently on the basis of the list of the teachers from the college prospectus and gets feedback from them regarding his publication. This method is less expensive and more simple but is unscientific and unreliable. This method results in more dependence on the enumerators. This method is appropriate for sample selection where the universe or population is not clearly defined or list of the units is not available or sample units are not clear in themselves.

(9) Extensive Sampling : In this method, sample size is taken almost as big as the population itself like 90% the section of the population. Only those units are left out for which data collection is

very difficult or almost impossible. Due to very large sample size, the method has greater level of accuracy. Intensive study of the problem becomes possible but this method involves heavy resources at disposal.

SAMPLING AND NON-SAMPLING ERRORS

The choice of a sample though may be made with utmost care, involves certain errors which may be classified into two types : (i) **Sampling Errors,** and (2) **Non-Sampling Errors.** These errors may occur in the collection, processing and analysis of data.

(1) Sampling Errors

Sampling errors are those which arise due to the method of sampling. Sampling errors arise primarly due to the following reasons:

(1) Faulty selection of the sampling method.
(2) Substituting one sample for the sample due to the difficulties in collecting the sample.
(3) Faulty demarcation of sampling units.
(4) Variability of the population which has different characteristics.

(2) Non-Sampling Errors

Non-sampling errors are those which creep in due to human factors which always varies from one investigator to another. These errors arise due to any of the following factors :

(1) Faulty planning.
(2) Faulty selection of the sample units.
(3) Lack of trained and experienced staff which collect the data.
(4) Negligence and non-response on the part of the respondent.
(5) Errors in compilation.
(6) Errors due to wrong statistical measures.
(7) Framing of a wrong questionnaire.
(8) Incomplete investigation of the sample survey.

Basic Concepts of Sampling

Sampling Distribution: The purpose of selecting and studying a sample from the population is to estimate or make inference about some population characteristics. In this process, the knowledge of the sampling distribution is of vital importance.

Some Important Terms:

The following terms are widely used in the study of the sampling distribution:

(1) Parameters: Any statistical measures computed from the population data is known as parameter. Thus, population mean, population standard deviation, population variance, population proportion, etc., are all parameters. Parameters are denoted by the Greek letters such as μ, σ^2, σ and P.

(2) Statistics: Any statistical measure computed from sample data is known as statistic. Thus, sample mean, sample standard deviation, sample variance, sample proportion, etc., are all statistics. Statistics are denoted by Roman letters such as $\overline{X}$, s, s^2 and p.

(3) Sampling with and without replacement: Sampling is a procedure of selecting a sample from the population. Sampling may be done with or without replacement. Sampling where each unit of a population may be chosen more than once is called sampling with replacement. If each unit cannot be chosen more than once, it is called sampling without replacement. In case of

sampling with replacement, the total number of possible samples each of size n drawn from a population of size N is N^n. But if the sampling is without replacement, the total number of possible samples will be $N_{c_n} = m$ (say).

Sampling Distribution of a Statistic

Sampling distribution constitutes the theoretical basis of statistical inference and is of considerable importance in business decision making. Sampling distribution of a statistic is the frequency distribution which is formed with various values of a statistic computed from different samples of the same size drawn from the same population. Suppose we draw all possible samples of size n from the population (N) with or without replacement. For each possible sample drawn from the population, we compute a statistic such as mean, median, standard deviation, variance, etc. The set of all possible values of a statistic is then classified and grouped into a frequency distribution (or probability distribution). The distribution so obtained is called the sampling distribution of a statistic. We could have various sampling distribution depending upon the nature of the statistic we have computed. If, for instance, the particular statistic computed is the sample mean, the distribution is called sampling distribution of mean. If, we compute variance of each sample, then it is called the sampling distribution of variance. Similarly, we could have sampling distributions of proportion, median, standard deviation, etc.

An Important Property of Sampling Distribution

An important property of the sampling distribution of a statistic is that if random samples of large size ($n > 30$) are taken from a population which may be normally distributed or not, then the sampling distribution of the statistic will approach a normal distribution.

Standard Error of a Statistic

The standard deviation of the sampling distribution of a statistic is known as the standard error of a statistic. As there are various types of sampling distributions, we could have various types of standard errors depending on the nature of sampling distribution. The standard deviation of the sampling distribution of means is called the standard error of the means. In sampling theory, instead of using the term standard deviation for measuring variation, we use a new term called standard error of mean. The standard error of mean measure the extent to which the sample mean differ from the population mean. Thus, **the basic difference between the standard deviation and standard error of mean is that the former measures the extent to which the individual items differ from the central value and the latter measures the extent to which individual sample mean differ from the population mean.** Like the standard error of the means, we could have standard error of the median, standard deviation, proportion, variance, etc.

Utility of Standard Error : The standard error is used in a large number of problems which are discussed as follows :

(1) Reliability of a Sample : The standard error gives an idea about the reliability and precision of a sample. That is, it indicates how much the estimated value differs from the observed values. The greater the standard error, the greater is the deviation between the estimated and observed values and lesser is the reliability of a sample. The smaller the standard error, the smaller is the deviation between the estimated and observed values and greater is the reliability of a sample.

(2) Tests of Significance : The standard error is also used to test the significance of the various results obtained from small and large samples. In case of large sample, if the difference between the observed and the expected value is greater than 1.96 standard error, then we reject the hypothesis

at 5% and conclude that sample differs widely from the population. But if the difference between the observed and the expected value is greater than 2.58 S.E. (Standard error), then we reject the null hypothesis at 1% and conclude that the sample differs widely from the population.

(3) To determine the confidence limits of the unknown population mean : The standard error enables us in determining the confidence limits within which a population parameter is expected to lie with a certain degree of confidence. The confidence limits of the unknown population mean μ are given by.

Large Sample	Small Sample
95% confidence limits for μ	95% confidence limit for μ
$\bar{x} - 1.96$ S.E. and $\bar{x} + 1.96$ S.E.	$\bar{x} \pm t_{.05}$ S.E.
99% confidence limits for μ	99% confidence limits for μ
$\bar{x} - 2.58$ S.E. and $\bar{x} + 2.58$ S.E.	$\bar{x} \pm t_{01}$ S.E.

SAMPLING DISTRIBUTION OF MEANS

It is an important sampling distribution widely used in the sampling theory. Draw all possible samples of size n with or without replacement from population of size N with mean μ and variance σ^2. For each possible sample drawn from the population, we compute the mean $\bar{x}$ of each sample. The mean will vary from sample to sample. The set of all possible means obtained from different samples is called the sampling distribution of means.

Properties : The following are the important properties of the sampling distribution of means:

(i) The mean of the sampling distribution of means is equal to the population mean (μ). Symbolically,

$$\mu_{\bar{x}} = \mu \qquad \text{or} \qquad E(\bar{x}) = \mu$$

This property can be proved as follows :

Let $x_1, x_2, \ldots, x_n$ represent a random sample (with replacement) of size n from a finite population of size N having its mean μ and variance σ^2, then

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

$$E(\bar{x}) = E\left[\frac{\Sigma x}{n}\right] = E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right]$$

$$= \frac{1}{n}\{E(x_1) + E(x_2) + \ldots + E(x_n)\}$$

$$= \frac{1}{n}\{\mu + \mu + \mu + \ldots + \mu\} = \frac{1}{n} \cdot n\,\mu = \mu$$

Thus, the mean of the sampling distribution of means is equal to the population mean.

(2) The standard error of the sampling distribution of means is obtained as :

$$S.E._{\bar{x}} \text{ or } \sigma_{\bar{x}} = \frac{\text{S.D. of Population}}{\sqrt{\text{Size of the sample}}} = \frac{\sigma}{\sqrt{n}}$$

This property can be proved as follows :

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{\Sigma x}{n}\right) = \text{Var}\left(\frac{x_1 + x_2 + \ldots x_n}{n}\right)$$

$$= \frac{1}{n^2}[\text{Var}(x_1) + \text{Var}(x_2) + \ldots \text{Var}(x_n)]$$

$$= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots \sigma^2]$$

$$= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

where, σ^2 is the population variance, x is the sample.

Because $n > 1$, obviously, $\frac{\sigma^2}{n} < \sigma^2 \Rightarrow V(\bar{x}) <$ Population variance.

$$\therefore \quad \text{S.E.}_{\bar{x}} = \sigma_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

This formula holds only when sampling is with replacement.

Note : When the population is finite and the samples are drawn without replacement, then $\text{S.E.}_{\bar{x}}$ is obtained as :

$$\text{S.E.}_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

(3) The sampling distribution of means is approximately a normal distribution with mean μ and variance σ^2, provided the sample is large ($n > 30$).

(4) The following formula is used to find the probability of the sampling distribution of means.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Let us illustrate the concept of sampling distribution of means by the following example :

Example 1. **Consider a population consisting of three values : 2, 5 and 8. Draw all possible samples of size 2 with replacement from the population. Construct sampling distribution of means. Also find the mean and standard error of the distribution.**

Solution. The population consists of three values.The total number of possible samples of size 2 drawn with replacement are $N^n = 3^2 = 9$. All possible random samples and their sample mean are shown is the following table :

Sample No.	Sample Values	Sample Mean $\bar{x}$
1.	(2, 2)	$\frac{1}{2}(2+2) = 2$
2.	(5, 2)	$\frac{1}{2}(5+2) = 3.5$
3.	(8, 2)	$\frac{1}{2}(8+2) = 5$
4.	(2, 5)	$\frac{1}{2}(2+5) = 3.5$
5.	(5, 5)	$\frac{1}{2}(5+5) = 5.0$
6.	(8, 5)	$\frac{1}{2}(8+5) = 6.5$

7.	(2, 8)	$\frac{1}{2}(2+8)=5.0$
8.	(5, 8)	$\frac{1}{2}(5+8)=6.5$
9.	(8, 8)	$\frac{1}{2}(8+8)=8.0$

On the basis of the means ($\bar{x}$) of all the 6 possible samples, the sampling distribution of means is given below :

Sample Means ($\bar{x}$)	f	$f\bar{x}$	$d=\bar{x}-\mu_{\bar{x}}$	d^2	fd^2
2	1	2	− 3	9	9
3.5	2	7	− 1.5	2.25	4.50
5.0	3	15	0	0	0
6.5	2	13	1.5	2.25	4.50
8.0	1	8	+ 3	9.0	9.0
	$\Sigma f = 9$	$\Sigma f\bar{x} = 45$			$\Sigma fd^2 = 27$

Mean of the Sampling Distribution of Means

$$\mu_{\bar{x}} = \frac{\Sigma f\bar{x}}{\Sigma f} = \frac{45}{9} = 5$$

Variance of the Sampling Distribution of Means

$$\text{Var}(\bar{x}) = \frac{\Sigma f(\bar{x}-\mu_{\bar{x}})^2}{\Sigma f} = \frac{\Sigma fd^2}{\Sigma f} = \frac{27}{9} = 3$$

Hence, $\text{S.E.}_{\bar{x}} = \sigma_{\bar{x}} = \sqrt{3} = 1.732$

Aliter : The sampling distribution of means can also be written in terms of probability as :

Sample Means ($\bar{x}$)	2	3.5	5.0	6.5	8.0
Probability (p)	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

Since 3.5 occurs twice, its probability of occurence is $\frac{2}{9}$, 5 occurs thrice, its probability of occurrence is $\frac{3}{9}$ and 6.5 occurrence is $\frac{2}{9}$. Each of the other sample mean occurs only once with probability $\frac{1}{9}$.

Mean of the Sampling Distribution of Means

$$E(\bar{x}) = \Sigma p\bar{x} = 2\times\frac{1}{9} + 3.5\times\frac{2}{9} + 5\times\frac{3}{9} + 6.5\times\frac{2}{9}\times 8.0\times\frac{1}{9}$$

$$= \frac{1}{9}\cdot[2+7+15+13+8] = \frac{45}{9} = 5$$

Variance of the Sampling Distribution of Means

$$\text{Var}(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$E(\bar{x}^2) = 2^2 \times \frac{1}{9} + 3.5^2 \times \frac{2}{9} + 5^2 \times \frac{3}{9} + 6.5^2 \times \frac{2}{9} + 8^2 \times \frac{1}{9}$$

$$= \frac{1}{9} \cdot [4 + 25 + 75 + 84.5 + 64]$$

$$= \frac{1}{9}[252.5] = 28.055$$

$$\text{Var}(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})^2] = 28.055 - 25 = 3.055 = 3$$

Hence, $\text{S.E.}_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \sqrt{3} = 1.732$

Example 2. **Construct a sampling distribution of the sample means from the following population :**

Population Unit :	1	2	3	4
Observation :	22	24	26	28

when random sample of size 2 are taken from it without replacement. Also find the mean and standard error of the distribution.

Solution. The population consists of four values (22, 24, 26, 28). The total number of possible sample of size 2 drawn without replacement are ${}^4C_2 = 6$. All the possible random samples and their sample means are shown in the table given below :

Sample No.	Sample Values	Sample Mean $\bar{x}$
1.	(22, 24)	$\frac{1}{2}(22 + 24) = 23$
2.	(22, 26)	$\frac{1}{2}(22 + 26) = 24$
3.	(22, 28)	$\frac{1}{2}(22 + 28) = 25$
4.	(24, 26)	$\frac{1}{2}(24 + 26) = 25$
5.	(24, 28)	$\frac{1}{2}(24 + 28) = 26$
6.	(26, 28)	$\frac{1}{2}(26 + 28) = 27$

On the basis of the means ($\bar{x}$) of all the 6 samples without replacement, the sampling distribution of mean is given below:

Sampling Distribution of Means without Replacement

Sample Means ($\bar{x}$)	f	$f\bar{x}$	$d = \bar{x} - \mu_{\bar{x}}$	d^2	fd^2
23	1	23	− 2	4	4
24	1	24	− 1	1	1
25	2	50	0	0	0
26	1	26	1	1	1
27	1	27	2	4	4
	$\Sigma f = 6$	$\Sigma f\bar{x} = 150$			$\Sigma fd^2 = 10$

Mean of the Sampling Distribution of Means

$$\mu_{\bar{x}} = \frac{\Sigma f\bar{x}}{\Sigma f} = \frac{150}{6} = 25$$

Variance of the Sampling Distribution of Means

$$\text{Var}(\bar{x}) = \frac{\Sigma f(\bar{x} - \mu_x)^2}{\Sigma f} = \frac{\Sigma fd^2}{\Sigma f} = \frac{10}{6} = \frac{5}{3}$$

Hence, $$S.E._{\bar{x}} = \sigma_{\bar{x}} = \sqrt{\text{Var}\,\bar{x}} = \sqrt{\frac{5}{3}} = 1.29$$

Aliter : The sampling distribution of means can also be written in terms of probability as below :

Sample Means ($\bar{x}$)	23	24	25	26	27
Probability (p)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Since, 25 occurs twice, its probability of occurence is $\frac{2}{6}$. Each of the other sample means occurs only once with probability $\frac{1}{6}$.

Mean of the Sampling Distribution of Means

$$E(\bar{x}) = \Sigma p\bar{x} = \frac{1}{6} \times 23 + \frac{1}{6} \times 24 + \frac{2}{6} \times 25 + \frac{1}{6} \times 26 + \frac{1}{6} \times 27$$

$$= \frac{1}{6} \cdot [23 + 24 + 50 + 26 + 27] = \frac{150}{6} = 25$$

Variance of the Sampling Distribution of Means

$$\text{Var}(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$E(\bar{x}^2) = 23^2 \times \frac{1}{6} + 24^2 \times \frac{1}{6} + 25^2 \times \frac{2}{6} + 26^2 \times \frac{1}{6} + 27^2 \times \frac{1}{6}$$

$$= \frac{1}{6} \cdot [529 + 576 + 1250 + 676 + 729]$$

$$= \frac{3760}{6} = 626.17$$

$$\text{Var}(\bar{x}) = E(\bar{x}^2) = [E(\bar{x})]^2 = 626.17 - 625 = 1.67$$

Hence, $$S.E._{\bar{x}} = \sigma_{\bar{x}} \sqrt{\text{Var}(\bar{x})} = \sqrt{1.67} = 1.29$$

Example 3. *A population consists of four elements : 3, 7, 11, 15. Consider all possible samples of size two which can be drawn with replacement from this population. Find (i) the population mean μ (ii) the population variance σ^2 (iii) the mean of the sampling distribution of means (iv) standard error (or S.D.) of the sampling distribution of means. Verify (iii) and (iv) by using (i) and (ii) and by use of suitable formula.*

Solution. (i) μ = population mean $= \frac{\Sigma X}{N} = \frac{3 + 7 + 11 + 15}{4} = \frac{36}{4} = 9$

(ii) σ^2 = population variance $= \dfrac{\Sigma (X-\mu)^2}{N} = \dfrac{(-6)^2+(-2)^2+(2)^2+(6)^2}{4} = \dfrac{80}{4} = 20$

$\therefore$ $\sigma = \text{S.D.} = \sqrt{20}$

(iii) All possible random samples of size two with replacement is $N^n = 4^2 = 16$ and their sample means are shown in the following table :

Sample No.	Sample Values	Sample Mean $\bar{x}$	Sample No.	Sample Values	Sample Mean $\bar{x}$
1.	(3, 3)	3	9	(11, 3)	7
2.	(3, 7)	5	10	(11, 7)	9
3.	(3, 11)	7	11	(11, 11)	11
4.	(3, 15)	9	12	(11, 15)	13
5.	(7, 3)	5	13	(15, 3)	9
6.	(7, 7)	7	14	(15, 7)	11
7.	(7, 11)	9	15	(15, 11)	13
8.	(7, 15)	11	16	(15, 15)	15

On the basis of the mean ($\bar{x}$) of all the 16 samples with replacement, the sampling distribution of $\bar{x}$ can be written as :

Sample Means ($\bar{x}$)	Frequency (f)	$f\bar{x}$	$d = \bar{x} - \mu_{\bar{x}}$	d^2	fd^2
3	1	3	− 6	36	36
5	2	10	− 4	16	32
7	3	21	− 2	4	12
9	4	36	0	0	0
11	3	33	+ 2	4	12
13	2	26	+ 4	16	32
15	1	15	+ 6	36	36
	$\Sigma f = 16$	$\Sigma f\bar{x} = 144$			$\Sigma fd^2 = 160$

Mean of the Sampling Distribution of Means

$$\mu_{\bar{x}} = \frac{\Sigma f\bar{x}}{\Sigma f} = \frac{144}{16} = 9$$

Variance of the Sampling Distribution of Means

$$\text{Var}(\bar{x}) = \frac{\Sigma f(\bar{x}-\mu_{\bar{x}})^2}{\Sigma f} = \frac{160}{16} = 10$$

Hence $$\text{S.E.}_{\bar{x}} = \sigma_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \sqrt{10}$$

Using the formula, $\mu_{\bar{x}} = \mu$ and $V(\bar{x}) = \dfrac{\sigma^2}{n}$, we get the mean of the sampling distribution of means $= \mu_{\bar{x}} = \mu = 9$ and variance of the sampling distribution of means $= \dfrac{\sigma^2}{n} = \dfrac{20}{2} = \dfrac{\sigma^2}{2} = 10.$

Hence, the results of (iii) and (iv) are verified by using the results of (i) and (ii).

Example 4. **A population consists of the following elements :**

2, 4, 5, 8, 11

Find:

(a) How many different samples of size 3 are possible when sampling is done without replacement.

(b) List all of the possible different samples.

(c) Compute the mean of each of the samples given in part (b).

(d) Find the sampling distribution of sample mean $\overline{X}$.

(e) If all the elements are euqally likely, compute the value of the population mean μ.

Solution. The population consists of five elements (2, 4, 5, 8, 11).

(a) The total number of possible samples of size 3 drawn without replacement are $5_{C_3} = 10$.

(b) All the possible different samples and their sample means are shown in the following table.

Sample No.	Sample Values	Sample Mean $\bar{x}$
1.	(2, 4, 5)	$\frac{1}{3}(2+4+5) = 3.67$
2.	(2, 4, 8)	$\frac{1}{3}(2+4+8) = 4.67$
3.	(2, 4, 11)	$\frac{1}{3}(2+4+11) = 5.67$
4.	(2, 5, 8)	$\frac{1}{3}(2+5+8) = 5.0$
5.	(2, 5, 11)	$\frac{1}{3}(2+5+11) = 6.0$
6.	(2, 8, 11)	$\frac{1}{3}(2+8+11) = 7.0$
7.	(4, 5, 8)	$\frac{1}{3}(4+5+8) = 5.67$
8.	(4, 5, 11)	$\frac{1}{3}(4+5+11) = 6.67$
9.	(4, 8, 11)	$\frac{1}{3}(4+8+11) = 7.67$
10.	(5, 8, 11)	$\frac{1}{3}(5+8+11) = 8.0$

In the above table, we have 10 possible samples of size 3 without replacement. Since, 5.67 occurs twice, its probability of occurrence is $\frac{2}{10}$. Each of the other sample means occur only once with probability $\frac{1}{10}$.

Sampling distribution of means $\bar{x}$ (i.e., the probability distribution of sample mean $\bar{x}$) is given below :

Sampling Distribution of $\bar{x}$

Sample Mean $\bar{x}$	3.67	4.67	5	5.67	6	6.67	6	7.67	8.0
Probability (p)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

(e) Population consists of the values (2, 4, 5, 8, 11). Since, each value occurs equally likely, the probability of occurence of each value is $\frac{1}{5}$. Hence,

Population Values (X) :	2	4	5	8	11
Probability (p) :	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

$$\text{Population Mean} = \mu = 2\times\frac{1}{5}+4\times\frac{1}{5}+5\times\frac{1}{5}+8\times\frac{1}{5}+11\times\frac{1}{5}$$

$$=\frac{1}{5}\cdot[2+4+5+8+11]=\frac{30}{5}=6$$

LAW OF LARGE NUMBERS AND CENTRAL LIMIT THEOREM

Law of Large Numbers and the Central Limit Theorem both serve the basis for the development of sampling distribution of a statistic.

Law of Large Numbers : The law of large numbers states that as the sample size increases, the sample mean would be closer and closer to the population mean. It does not gurantee that if the sample size is increased sufficiently, the sample mean will be equal to the population mean. There are two implications of the law of large numbers (i) the difference between sample mean and population mean can be reduced by increasing the sample size, and (ii) variation from one sample mean to another sample mean (of the same size) also decreases as the size of the sample increases.

Central Limit Theorem : It is widely used in the field of estimation and inference. This theorem states that if we select random sample of large size n from any population with mean μ and standard deviation σ and compute the mean of each sample, then the sampling distribution of mean $\bar{x}$ approaches normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. This is true even if the population itself is not normal. The utility of this theorem is that it requires virtualy no conditions on the distribution pattern of the population.

QUESTIONS

1. Distinguish between population and sample. Discuss the relative merits and demerits of census and sampling Methods.
2. Explain the various methods of sampling. Also discuss their relative merits and demerits.
3. Explain the simple random sampling technique and state when it is used.
4. (a) What is a random sample ? Discuss the various methods of drawing a random sample.
   (b) Distinguish between sampling and non-sampling errors.

5. What is meant by sampling distribution of a statistic? Also, define standard error of a statistic.

OR

Discuss briefly the concept of sampling distribution of an estimator.

6. Distinguish between :

(i) Population and Sample (ii) Parameters and Statistics

(iii) Sampling with and without replacement.

7. Write a note on "Sampling Distribution of Means."

8. A population consists of five numbers (2, 3, 6, 8, 11). Draw all possible random samples of size 2 which can be drawn with replacement from this population. Construct a sampling distribution of means. Also, find the mean and standard error of the distribution.

9. A population consists of four numbers (3, 7, 11, 15). Consider all possible samples of size 2 which can be drawn without replacement from this population. Construct the sampling distribution of means. Also find the mean and standard error of the distribution.

10. A population consists of the following five elements :

3, 5, 9, 11, 17

Find :

(a) How many different samples of size 3 are possible when sampling is done without replacement ?

(b) List all of the possible different samples.

(c) Compute the sample mean for each of the samples given in part (b).

(d) Find the sampling distribution of the sample mean $\bar{x}$. Use a probability histogram to graph the sampling distribution of $\bar{x}$.

(e) If all five population values are equally likely, compute the value of the population mean, μ.

11. A population consists of numbers : 2, 3, 6, 8, 11. (i) Enumerate all possible samples of size 2 which can be drawn from this population (without replacement) (ii) Calculate the mean of the sampling distribution of means and show that the mean of the sampling distribution of means is equal to the population mean. (iii) Calculate the variance of the sampling distribution of means and show that it is less than that population variance.

12. (a) Show that the mean of the sampling distribution of means is equal to the population mean *i.e.*, $E(\bar{x}) = \mu$.

(b) Derive the variance of the sampling distribution of the sample mean. Is it more than population variance ?

13. Give statements of Law of Large Numbers and Central Limit Theorem. Discuss their significance in sampling theory.

❀❀❀

3

Tests of Hypothesis – Large Sample Tests

INTRODUCTION

The main objective of the sampling theory is the study of the **Tests of Hypothesis or Tests of Significance.** In many circumstances, we are to make decisions about the population on the basis of only sample information. For example, on the basis of sample data, (i) a quality control manager is to determine whether a process is working properly, (ii) a drug chemist is to decide whether a new drug is really effective in curing a disease, (iii) a statistician has to decide whether a given coin is unbiased, etc. Such decisions are called statistical decisions (a simply decisions). The theory of testing of hypothesis employs various statistical techniques to arrive at such decisions on the basis of the sample study.

BASIC CONCEPTS OF HYPOTHESIS TESTING

The following basic concepts are used in the study of tests of hypothesis:

(1) Hypothesis (or Statistical Hypothesis) : In attempting to arrive at decisions about the population on the basis of sample information, it is necessary to make assumptions about the population parameters involved. Such an assumption (or statement) is called a statistical hypothesis which may or may not be true.

There are two types of hypothesis:

(a) Null Hypothesis and (b) Alternative Hypothesis.

(a) Null Hypothesis : In tests of hypothesis we always begin with an assumption or hypothesis (*i.e.*, assumed value of a population parameter). This is called Null Hypothesis. The null hypothesis asserts that there is no (significant) difference between the sample statistic and the population parameter and whatever the observed difference is there, is merely due to fluctuations in sampling from the same population. Null hypothesis is usually denoted by the symbol H_0. **R.A. Fisher** defined null hypothesis as "the hypothesis which is tested for possible rejection under the assumption that it is true". In other words, the hypothesis (regarding some characteristic of population) which is to be verified with the help of a random sample or the hypothesis which is under test is called null hypothesis. For example, if we want to test the hypothesis that the mean of the population to be taken as μ_0, then the null hypothesis (H_0) is $\mu = \mu_0$.

(b) Alternative Hypothesis : Any hypothesis different from the null hypothesis (H_0) is called an alternative hypothesis and is denoted by the symbol H_1. The two hypothesis H_0 and H_1 are such that if one is accepted, the other is rejected and vice versa. For example, if we want to test whether the population mean μ has a specified value μ_0, then (i) Null Hypothesis is $H_0 : \mu = \mu_0$ and

(ii) Alternative Hypothesis may be (a) $H_1 : \mu \neq \mu_0$ (*i.e.*, $\mu > \mu_0$ or $\mu < \mu_0$), or (b) $H_1 : \mu > \mu_0$ or (c) $H_1 : \mu < \mu_0$. Thus, there can be more than one alternative hypothesis.

(2) Type I and Type II Errors : In the process of hypothesis testing we usually come across same sort of errors, called **errors in hypothesis testing** which are grouped in two types as:

(i) Type I Errors and (ii) Type II Errors.

(i) Type I Errors : Type I errors are made when we reject the null hypothesis H_0 though it is true. In other words, when H_0 is rejected despite its being true, then it is called Type I errors. The probability of making a type I error is denoted by $p(E_1) = \alpha$ and the probability of making a correct decision is then $1 - \alpha$ *i.e.*, $p = 1 - \alpha$.

(ii) Type II Errors : Type II errors are made when we accept the null hypothesis though it is false. In other words, when H_0 is accepted despite its being false, then it s called Type II errors. The probability of making a type II error is denoted by β. Thus, $p(E_2) = \beta$.

The following table illustrates Type I and Type II errors.

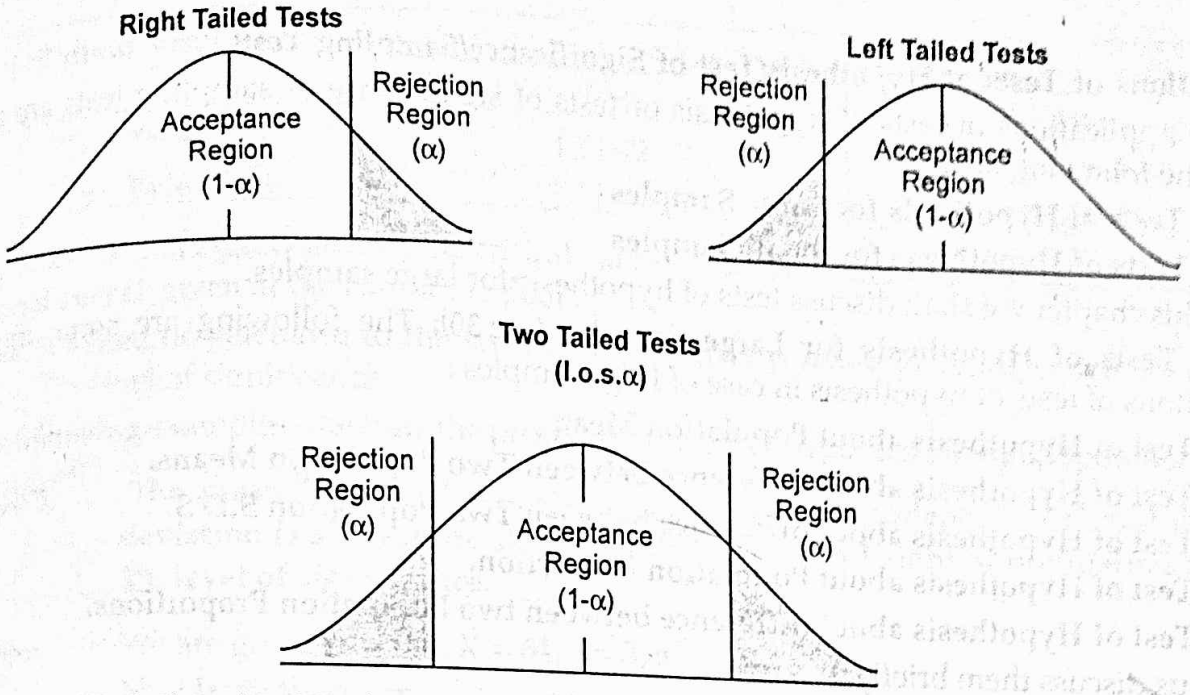
	To Accept H_0	To Reject H_0
H_0 is True	Correct Decision $p = 1 - \alpha$	Type I Error, $p = \alpha$
H_0 is False	Type II Error $p = \beta$	Correct Decision $p = 1 - \beta$

While testing hypothesis, attempts are made to minimise both the types of errors, although it is not at all possible to reduce them both at the same time.

(3) Level of Significance : This refers to the degree of significance with which we accept or reject a particular hypothesis. Since 100% accuracy is not possible in taking a decision over the acceptance or rejection of a hypothesis, we have to take the decision at a particular level of confidence which would speak of the probability of one being correct or wrong in accepting or rejecting a hypothesis. In most of the cases of hypothesis testing, such a confidence is fixed at 5% level, which implies that our decision would be correct to the extent of 95%. For a greater precise, however, such a confidence may be fixed at 1% level which would imply that the decision would be correct to the extent of 99%. This level is usually denoted by the symbol, α (alpha) which represents the probability of committing the type I error (*i.e.* rejecting a null hypothesis which is true). The level of confidence (or significance), is always fixed in advance before applying the test procedures. It is important to note that if no level of significance is given, then we always take $\alpha = 0.05$.

(4) Critical Region or Rejection Region : The critical region or rejection region is the region of the standard normal curve corresponding to a pre-determined level of significance. The region under the normal curve which is not covered by the rejection region is known as Acceptance Region. Thus, the statistic which leads to the rejection of null hypothesis H_0 gives us a region known as Rejection Region or Critical Region. While those which lead to acceptance of H_0 give us a region called as Acceptance Region.

(5) One Tailed Test and Two Tailed Test : A test of any statistical hypothesis where the alternative hypothesis is expressed by the symbol (<) or the symbol (>) is called a one tailed test since the entire critical region lies in one tail of the distribution of the test statistic. The critical region for all alternative hypothesis containing the symbol (>) lies entirely on the right tail of the distribution while the critical region for an alternative hypothesis containing a less than (<) symbol lies entirely in the left tail. The symbol indicates the direction where the critical region lies. A test of any statistical hypothesis where the alternative is written with a symbol '$\neq$' is called a two-tailed



test, since the critical region is split into two equal parts, one in each tail of the distribution of the test statistic. The following figures illustrate one tailed and two tailed tests :

(6) **Critical Value :** The critical values of the standard normal variate (Z) for both the two-tailed and one tailed tests at different level of significance are very often required in hypothesis testing. The following table gives critical values for both one tailed and two tailed tests at various level of significance.

Level of Significance (α)	α = 0·10	α = 0·05	α = ·01	α = ·005
Critical values of Z (for one tailed test)	– 1·28 or + 1·28	– 1·645 or + 1·645	– 2·33 or + 2·33	– 2·58 or + 2·58
Critical values of Z (for two tailed test)	– 1·645 or + 1·645	– 1·96 or + 1·96	– 2·58 or + 2·58	– 2·81 or + 2·81

Note : For other level of significance α, the critical value of Z can be found from the table **"Area under the Normal Curve."**

Procedure of Testing a Hypothesis:

Testing of a hypothesis passes through the following steps :

(1) **Set up a null hypothesis :** It is denoted H_0. Null hypothesis assumes that difference between any values to be compared is not significant.

(2) **Set up a suitable level of significance :** A suitable level of significance is determined to test the null hypothesis. In practice, 5% significance level is used.

(3) **Set up a suitable test of statistic :** A number of test statistics like Z, t, χ^2, F etc. may be applied to test the null hypothesis. It is decided only on the basis of available information.

(4) **Doing necessary calculation :** After selecting appropriate statistic, computations relating to the test statistic are made and values are worked out.

(5) **Making Decisions :** In the process of hypothesis testing, results are interpreted at the final stage. For this purpose, we compare the computed value of a test statistic with the table value at a pre-determined level of significance. If computed value is greater that the table value at 5% or 1% level of significance, then null hypothesis is rejected. In such a situation, the sample does not represent population.

Applications of Tests of Hypothesis/Test of Significance/Sampling Tests

The applications of tests of hypothesis or tests of significance or sampling tests are studied under the following heads :

(A) Tests of Hypothesis for Large Samples

(B) Tests of Hypothesis for Small Samples

In this chapter we shall discuss tests of hypothesis for large samples.

(A) Tests of Hypothesis for Large Samples ($n \geq 30$): The following are some important applications of tests of hypothesis in case of large samples :

(1) Test of Hypothesis about Population Mean

(2) Test of Hypothesis about difference between Two Population Means.

(3) Test of Hypothesis about difference between Two Population S.D'S.

(4) Test of Hypothesis about Population Proportion.

(5) Test of Hypothesis about difference between two Population Proportions.

Let us discuss them briefly.

(1) Test of Hypothesis about Population Mean μ : The test of hypothesis concerning population mean μ in case of large sample requires the use of normal distribution. Let $\overline{X}$ be the mean of a large random sample of size n drawn from a normal population with mean μ and standard deviation σ. To test the hypothesis that population mean μ has a specified value, the appropriate test statistic to be used is :

$$Z = \frac{\overline{X} - \mu}{S.E_{\overline{X}}}$$

Where, $\overline{X}$ = sample mean; μ = population mean, $S.E_{\overline{X}}$ = Standard Error of Mean.

Procedure : The following steps are taken for test of hypothesis about population mean:

(i) Set up the null hypothesis $H_0 : \mu = \mu_0$ *i.e.*, there is no difference between the sample mean and population mean. Alternative Hypothesis : $H_1 : \mu \neq \mu_0$ (Two tailed test).

or $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ (One tailed test)

(ii) Compute the $S.E_{\overline{X}}$ by using the following formula :

(a) $S.E_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ When population S.D. σ is known.

(b) $S.E_{\overline{X}} = \frac{s}{\sqrt{n}}$ When sample S.D. is known.

(iii) Substituting the value of $\overline{X}$, μ and $S.E_{\overline{X}}$ in the above stated Z-statistic.

(iv) Select the desired level of significance α if not given and corresponding to that level of significance (l.o.s.), we find the critical value of Z_α from the table **"Areas under the normal curve".**

(v) The computed value of Z is compared with the critical value of Z. If the computed value of $Z = |Z| <$ critical value of Z at a level of significance α, then we accept the null hypothesis H_0 and if the computed value of $Z = |Z| >$ critical value of Z at a level of significance α, then we reject the null hypothesis and accept the alternative hypothesis H_1.

Note :

1. *If the population S.D. σ is not known, then sample S.D. (s) is used for large samples.*
2. *The critical value of Z_α (for large samples) corresponding to various level of significance are given below :*

Critical Value (Z_α)	Level of Significance (α)	
	1%	5%
Two Tailed Test	$\|Z\| = 2.58$	$\|Z\| = 1.96$
One Tailed Test	$\|Z\| = 2.33$	$\|Z\| = 1.64$

For other level of significance α, the critical values Z can be found from the table **"Area under the normal curve"** given at the end of the book.

Note 3. When no reference to the level of significance is made given, then we always take $\alpha = .05$ *i.e.*, 5% level of significance.

The following examples illustrate the procedure for test of hypothesis about population mean :

Example 1. **The mean height of a random sample of 100 students is 64″ and standard deviation is 3″. Test the statement that the mean height of population is 67″ at 5% level of significance.**

Solution. We are given : $n = 100, \bar{X} = 64, s = 3, \mu = 67$

Null Hypothesis :

$$H_0 : \mu = 67$$

Alternative hypothesis : $H_1 : u \neq 67$ (⇒ Two tailed test)

$$S.E._{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{3}{\sqrt{100}} = \frac{3}{10} = 0.3 \quad \text{(For large samples, } \sigma = s\text{)}$$

Now, we compute Z-statistic as :

$$|Z| = \frac{|\bar{X} - \mu|}{S.E._{\bar{X}}} = \frac{64 - 67}{0.3} = 10$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of $|Z|$ is more than the critical value of Z at 5% level of significance, we reject the null hypothesis and conclude that the population mean cannot be equal 67.

Example 2. **A random of 400 male students is found to have a mean height of 171·38 cm. Can it be reasonably regarded as a sample from a large population with mean height 171·17 cm and standard deviation 3·30 cm ? Use $\alpha = .05$.**

Solution. We are given : $n = 400, \bar{X} = 171.38, \mu = 171.17, \sigma = 3.30$

Null Hypothesis $H_0 : \mu = 171.17$ *i.e.*, the sample has been drawn from population with $\mu = 171.17$ and $\sigma = 3.30$.

Alternative Hypothesis $H_1 : \mu \neq 171.17$ (⇒ Two tailed test)

$$S.E._{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.30}{\sqrt{400}} = 0.165$$

$$|Z| = \frac{\bar{X} - \mu}{S.E._{\bar{X}}} = \frac{171.38 - 171.17}{0.165} = 1.273$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of $|Z|$ is less than the critical value of Z at 5% level of significance, we accept the null hypothesis and conclude that the population mean

is equal to 171·17 *i.e.*, the sample has been drawn from the population with mean 171·17.

Example 3. **A stenographer claims that she can take dictation at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trials in which she demonstrates a mean of 116 words with a standard deviation of 15 words. (Use 5% l.o.s.).**

Solution. **We are given :** $n = 100, \bar{X} = 116, s = 15, \mu = 120$

Null hypothesis $H_0 : \mu = 120$ (*i.e.*, the claim is accepted)

Alternative hypothesis $H_1 : \mu < 120$ (*i.e.*, the claim is rejected)

$$S.E._{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

$$|Z| = \frac{|\bar{X} - \mu|}{S.E._{\bar{X}}} = \frac{|116 - 120|}{1{\cdot}5} = \frac{4}{1{\cdot}5} = 2{\cdot}6$$

At 5% l.o.s., the critical value of Z for one tailed test = 1·645.

Since, the calculated value of $Z > 1{\cdot}645$, we reject H_0 and conclude that the claim of the stenographer is rejected.

Aliter : This question can also be solved by using two tailed test. Let us have; $H_0 : \mu = 120$ and $H_1 : \mu \neq 120$, it is a two tailed test.

And $H_1 : \mu \neq 120$, it is a two tailed test.

$$|Z| = 2.6$$

At 5% l.o.s. $Z_{.05} = 1.96$ (for two tailed test)

Since $|Z| > 1.96$, we reject H_0 and conclude that the claim of the stenographer is rejected.

Example 4. **An educator claims that the average IQ of government college students is no more than 110. To test this claim, a random sample of 150 students was taken and given relevant tests. Their average IQ score come to 111.2 with a standard deviation of 7·2. At level of significance of 0.01, test if the claim of the educator is justified.**

Solution. We are given: $n = 150, \bar{X} = 111.2, s = 7.2, \mu = 110$

Null hypothesis $H_0 : \mu \leq 110$

Alternative hypothesis : $\mu > 110$ (Right tailed test)

$$S.E_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{7{\cdot}2}{\sqrt{150}} = \frac{7{\cdot}2}{12{\cdot}25} = 0{\cdot}59.$$

$$|Z| = \frac{|\bar{X} - \mu|}{S.E_{\bar{X}}} = \frac{111{\cdot}2 - 110}{0{\cdot}59} = \frac{1{\cdot}2}{.59} = 2{\cdot}03$$

At 1% level of significance, the critical value of Z for one tailed test 2·33.

Since, the calculated value of $|Z|$ < critical value of Z at 1%, we accept H_0. This means that the claim of the educator is justified.

EXERCISE – 1

1. A random sample of 100 students gave a mean weight of 58 kg with S.D. of 4 kg. Test the hypothesis that the mean weight of the population is 60 kg. [Ans. $|Z| = 5$, H_0 is rejected]
2. A sample of 100 units is found to have 5 lbs as mean. Could it be regarded as a simple random sample from a large population whose mean is 5.64 lbs and $\sigma = 1{\cdot}5$ lbs. Use $\alpha = {\cdot}05$. [Ans. $|Z| = 4{\cdot}26$, H_0 is rejected]
3. The manufacturer of a particular make of a small car claim that on an average the car is driven 2000 kms per month. A random sample of 100 owners of the car are asked to keep a record of kilometers they drive their cars. On the basis of these sample records, it was found that an average the car was driven 2200 kms per month with a standard deviation of 600 kms. Do the sample data support the hypothesis that the average distance the car is driven has increased ? Use $\alpha = {\cdot}05$. [Ans. $Z = 3{\cdot}33$, H_0 is rejected]
4. A company claims that life of its product is 1600 hours. A sample of 100 units was tested and mean life of its products was found to be 1570 hours with a standard deviation of 120 hours. Test the claim of the company at 5% level of significance ? [Ans. $|Z| = 2{\cdot}5$, H_0 is rejected]
5. A weighing machine without any display was used by an average of 320 persons a day with a standard deviation of 50 parsons. When an attractive display was used on the machine, the average for 100 days increased by 15 persons. Can we say that the display did not keep much ? Use a level of significance of $0{\cdot}05$. [Ans. $Z = 3$, rejected H_0]

---

(2) Test of Hypothesis about difference two Population Means : Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of two independent random samples of large sizes n_1 and n_2 drawn from two populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2. To test whether the two population means are equal or not, the appropriate test statistic to be used is :

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S.E._{\bar{X}_1 - \bar{X}_2}}$$

Where, $\bar{X}_1$ = Mean of 1st Sample, $\bar{X}_2$ = Mean of 2nd Sample $S.E._{\bar{X}_1 - \bar{X}_2}$ = Standard error of the difference of two means.

Procedure : The following steps are taken for test of hypothesis about difference between two populations means :

(i) Set up the null hypothesis $H_0 : \mu_1 = \mu_2 = 0$ *i.e.*, $\mu_1 = \mu_2$ there is no difference between the two population means.

Alterantive Hypothesis : $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

or $H_1 : \mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (One tailed test)

(ii) Compute the $S.E._{\bar{X}_1 - \bar{X}_2}$ by using the following formulae :

(a) $S.E._{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ When Population S.D.s σ_1 and σ_2 are given

(b) $S.E._{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ When S.D. s_1 and s_2 of the two samples are given.

(c) $S.E._{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ When two random samples have been drawn form the same population with S.D. σ.

(d) $S.E._{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}-2r\cdot\frac{s_1\times s_2}{n_1\times n_2}}$ When two samples are correlated.

(iii) Substituting the values of $\bar{X}_1, \bar{X}_2, \mu_1, \mu_2$ and $S.E._{\bar{X}_1-\bar{X}_2}$ in the above stated Z-statistics.

The other steps such as (i) Level of significance, (ii) Critical Value of Z_α; (iii) Decision making for testing hypothesis of the difference between two means are the same as those given in test of hypothesis about population mean μ.

Example 5. **A random sample of 1000 workers from South India show that their mean wages are Rs. 47 per week with a standard deviation of Rs. 28. A random sample of 1500 workers from North India gives a mean wages of Rs. 49 per week with a standard deviation of Rs. 40. Is there any significant difference between the mean level of wages in two places ?**

Solution. We are given : $n_1 = 1000, \bar{X}_1 = 47, s_1 = 28$

$n_2 = 1500, \bar{X}_2 = 49, s_2 = 40$

Null Hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, there is no significant difference between two mean wages.

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$S.E._{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{(28)^2}{1000}+\frac{(40)^2}{1500}} = 1\cdot 36$$

$$|Z| = \frac{|\bar{X}_1-\bar{X}_2|}{S.E._{\bar{X}_1-\bar{X}_2}} = \frac{|47-49|}{1\cdot 36} = \frac{2}{1\cdot 36} = 1\cdot 47$$

At 5% level of significance, the critical value of Z for two tailed test $= \pm 1\cdot 96$.

Since, the calculated value of $|Z|$ is less than the critical value of Z, we accept the null hypothesis and concluded that there is no significant difference between the two mean level of wages.

Example 6. **The mean yield of wheat from Patiala District was 210 kgs with a standard deviation 10 kgs per acre from a sample of 100 plots. In another district Ludhiana, the mean yield was 220 kgs with standard deviation 12 kgs from a sample of 150 plots. Assuming that the standard deviation of the yield in the entire state was 11 kgs, test whether there is any significant difference between the mean yield of crops in the two districts.**

Solution. We are given : $n_1 = 100, \bar{X}_1 = 210, s_1 = 10$

$n_2 = 150, \bar{X}_2 = 220, s_2 = 12$

S.D. of population $= \sigma = 11$

Null hypothesis : $H_0 : \mu_1 = \mu_2$ *i.e.*, there is no significant difference between the mean yield of crops in two districts.

Alternative hypothesis : $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$S.E._{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(11)^2\left(\frac{1}{100} + \frac{1}{150}\right)}$$

$$= \sqrt{\frac{121}{100} + \frac{121}{150}} = \sqrt{1 \cdot 21 + 0 \cdot 807} = \sqrt{2 \cdot 017} = 1 \cdot 42$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{S.E._{\bar{X}_1 - \bar{X}_2}} = \frac{|210 - 220|}{1 \cdot 42} = 7 \cdot 04$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of $|Z| = 7 \cdot 05 >$ critical value of Z, we reject H_0 in favour of H_1. It means that there is a significant difference between the mean yield of crops in two districts.

ample 7. **Following information is available in respect of two brands of bulbs (Price same):**

	Brand A	Brand B
Mean Life (Hrs.)	1300	1248
S.D. (Hrs.)	82	93
Sample Size	100	100

Which brand should be preferred at 5 percent level of significance ?

olution. We are given : $n_1 = 100, \bar{X}_1 = 1300, s_1 = 82$

$n_2 = 100, \bar{X}_2 = 1248, s_2 = 93$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, there is no significant difference in the mean life of the two brands of bulb.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$S.E._{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(82)^2}{100} + \frac{(93)^2}{100}}$$

$$= \sqrt{\frac{6724}{100} + \frac{8649}{100}} = \sqrt{67 \cdot 24 + 86 \cdot 49} = 12 \cdot 399$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{S.E._{\bar{X}_1 - \bar{X}_2}} = \frac{1300 - 1248}{12 \cdot 399} = 4 \cdot 19$$

At 5% level, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of Z > the critical value of Z, we reject the hypothesis and conclude that there is a significant difference in the mean life of the two brands of bulbs. We should prefer to buy the bulbs of branch A since its average life is more.

Example 8. **A sample of heights of 6400 soldiers has a mean of 67·85 inches and a standard deviation of 2·56 inches while a sample of heights of 1600 sailors has a mean of 68·55 inches and a standard deviation of 2·52 inches. Do the data indicate that the sailors are on the average taller than the soldiers ?**

Solution. We are given : $n_1 = 6400, \bar{X}_1 = 67 \cdot 85, s_1 = 2 \cdot 56$

$n_2 = 1600, \bar{X}_2 = 68 \cdot 55, s_2 = 2 \cdot 52$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, there is no significant difference in the mean height of soldiers and sailors.

Alternative hypothesis $H_1 : \mu_2 > \mu_1$ or $\mu_1 < \mu_2$ ($\Rightarrow$ one tailed test)

$$\text{S.E.}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(2 \cdot 56)^2}{6400} + \frac{(2 \cdot 52)^2}{1600}} = 0 \cdot 071$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{S.E.}_{\bar{X}_1 - \bar{X}_2}} = \frac{|67 \cdot 85 - 68 \cdot 55|}{0 \cdot 071} = \frac{|-0 \cdot 7|}{0 \cdot 071} = 9 \cdot 859$$

At 5% level, the critical value of Z for one tailed test = 1·645.

Since, the calculated value $|Z|$ > critical value of Z at 5% level, we reject H_0 in favour of H_1. It means that the data indicate that the sailors are on an average taller than the soldiers.

Example 9. **The means of two large sample of sizes 1000 and 2000 are 168·75 cms and 170 cms respectively. Can the samples be regarded as drawn from a population with same mean and S.D. 6·25 cms.**

Solution. We are given : $n_1 = 1000, \bar{X}_1 = 168 \cdot 75$

$n_2 = 2000, \bar{X}_2 = 170$

$\sigma = 6 \cdot 25$

Null Hypothesis : $H_0 : \mu_1 = \mu_2$ *i.e.*, both the samples are drawn from the population with mean and S.D. 6·25.

Alternative Hypothesis : $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$\text{S.E.}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$= \sqrt{(6 \cdot 25)^2 \left(\frac{1}{1000} + \frac{1}{2000} \right)}$$

$$= \sqrt{0 \cdot 03906 + 0 \cdot 01953} = 0 \cdot 242$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{S.E.}_{\bar{X}_1 - \bar{X}_2}} = \frac{|168 \cdot 72 - 170|}{0 \cdot 242} = \frac{1 \cdot 28}{0 \cdot 242} = 5 \cdot 28$$

At 5% level, the critical value of Z for two tailed test = 1·96.

Since the calculated value of $|Z| = 5 \cdot 28 >$ critical value of $Z = 1 \cdot 96$, we reject H_0 and conclude that both the samples are not drawn from the population with same mean and S.D. 6·25.

Example 10. **Two different samples from two districts yielded the following results :**

District A : $\bar{X}_1 = 648, \quad s_1^2 = 120, \quad n_1 = 100$

District B : $\bar{X}_2 = 495, \quad s_2^2 = 140, \quad n_2 = 90$

Test at 0·05 level of significance that $\mu_1 - \mu_2 > 150$.

Solution. We are given : $\bar{X}_1 = 648, \quad s_1^2 = 120, \quad n_1 = 100$

$\bar{X}_2 = 495, \quad s_2^2 = 140, \quad n_2 = 90$

Null hypothesis $H_0 : \mu_1 - \mu_2 = 150$.

Alternative hypothesis $H_1 : \mu_1 - \mu_2 > 150$ ($\Rightarrow$ Right tailed test)

$$\text{S.E.}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= \sqrt{\frac{120}{100} + \frac{140}{90}} = \sqrt{1{\cdot}20 + 1{\cdot}55}$$

$$= \sqrt{2{\cdot}75} = 1{\cdot}658$$

$$|Z| = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{S.E.}_{\bar{X}_1 - \bar{X}_2}}$$

$$= \frac{(648 - 495) - (150)}{1{\cdot}658} = \frac{153 - 150}{1{\cdot}658}$$

$$= \frac{3}{1{\cdot}658} = 1{\cdot}809$$

At 5% level of significance, the critical value Z for one tailed test $= 1{\cdot}645$.

Since, the calculated value of $|Z|$ > the critical value of Z, we reject H_0 in favour of H_1. It means that the difference of the two means is greater than 150. *i.e.*, $\mu_1 - \mu_2 > 150$.

Example 10A. **In an intelligence test administrated to 60 fathers and them 100 children, the following results were obtained :**

Father's mean score 114; standard deviation 13

Son's mean score 110; standard deviation 11

Assuming the coefficient of correlation between them is + 0·75, calculate the standard error of the two means and state whether the difference is significant ?

Solution. $H_0 : \mu_1 = \mu_2$ *i.e.*, there is no significant difference in the mean score.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

We are given : $n_1 = 60, \quad \bar{X}_1 = 114; \quad s_1 = 13$

$n_2 = 100, \quad \bar{X}_2 = 110; \quad s_2 = 11$

$r = +{\cdot}75$

$$\text{S.E.}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2r \cdot \frac{s_1 \times s_2}{n_1 \times n_2}}$$

$$= \sqrt{\frac{(13)^2}{60} + \frac{(11)^2}{100} - 2 \times 0{\cdot}75 \times \frac{13 \times 11}{60 \times 100}}$$

$$= 2$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{S.E._{\bar{X}_1 - \bar{X}_2}} = \frac{|114 - 100|}{2}$$

$$= 7$$

At 5% level of significance, the critical value of $Z = 1{\cdot}96$ for two tailed test. Since the calculated value of Z is greater than the critical value of Z, we reject H_0 in favour of H_1. It means that there is significant difference in the mean scores of fathers and their sons.

EXERCISE – 2

1. The following information relates to wages of workers of two factories A and B. Test whether there is any significant difference between their mean wages. Use $\alpha = {\cdot}05$.

	Factory A	Factory B
Mean Wages (Rs.) :	100	105
Standard Deviation :	16	24
No. of Workers :	800	1600

[**Ans.** $Z = 6{\cdot}061$, H_0 is rejected]

2. A researcher claims that American 18 year old females are, on an average, taller than the British 18 year old females. To test this claim, a random sample of 50 American females and 50 British females was taken and their measurments are summaried as follows :

	American	British
Average height (in inches)	$\bar{X}_1 = 65.2$	$\bar{X}_2 = 64.5$
Standard deviation	$s_1 = 2.5$	$s_2 = 2.8$

Test the hypothesis that American females are taller than their British Counterparts at $\alpha = 0.05$

Hint : Use one tailed test. [**Ans.** $|Z| = 1.32$, Accept H_0]

3. A man buys 200 electric bulbs of each of 'Philips' and 'HMT'. He finds that Philips bulbs has a mean life of 2560 hours and a S.D. of 80 hours and HMT bulbs has a mean life of 2650 hours with a S.D. of 75 hours. Is there is a significant difference in the mean life of these two kinds of bulbs ? [**Ans.** $|Z| = 11{\cdot}612$, H_0 is rejected]

4. Two different samples from two districts yielded the following results :

District A : $\bar{X}_1 = 13{,}000$, $s_1 = 1300$, $n_1 = 100$

District B : $\bar{X}_2 = 13{,}900$, $s_2 = 1400$, $n_2 = 200$

Test at $0{\cdot}05$ level of significance that $\mu_1 - \mu_2 > 500$. [**Ans.** $Z = 2{\cdot}448$, H_0 is rejected]

5. 490 boys and 450 girls appeared at an examination in commerce. The mean and standard deviation of marks of boys are $54{\cdot}3$ and $17{\cdot}5$ respectively, whereas those of girls are $50{\cdot}6$ and $18{\cdot}0$. Is there a significant difference in marks of boys and girls at 1% level ?

[**Ans.** $Z = 3{\cdot}18$, H_0 is rejected]

6. 100 students of a college were put to tests in statistics and Accountancy respectively and the following results were obtained :

   Mean marks in Statistics = 45; S. D. = 7

   Mean marks in Accountancy = 43 ; S. D. = 6

   between marks in the two subjects = + 0·75.

   Calculate the standard error of the difference of the two means and state whether the difference is significant. [Ans. $S.E._{\bar{X}_1 - \bar{X}_2} = 0{\cdot}92$, $Z = 2{\cdot}17$, Reject H_0]

7. A test in statistics was conducted for a class of 70 boys and 60 girls. The test provides the following information :

   Boys : $n_1 = 70,\ \bar{X}1 = 70,\ \Sigma(X_1 - \bar{X}1)^2 = 7500$

   Girls : $n_2 = 60,\ \bar{X}2 = 65,\ \Sigma(X_2 - \bar{X}2)^2 = 7800$

   Test whether there is a significant difference between the performance of boys and girls at 5% level of significance. [Ans. $|Z| = 2.60$, reject H_0]

8. Intelligence test of two groups of boys and girls gave the following results :

	Mean	S.D.	N
Girls :	61	2	84
Boys :	60	4	100

   Is there a significance difference in the mean score obtained by boys and girls.
   Use $\alpha = .05$ [Ans. $Z = 2.12$, reject H_0]

---

(3) Test of Hypothesis about difference between two population standard deviations : Let s_1 and s_2 be the standard deviation of two independent random samples of sizes n_1 and n_2 from two populations wtih standard deviations σ_1 and σ_2 respectively. To test whether the two population S.D.'s are equal or not, one appropriate test statistic is to be used as

$$Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E._{s_1 - s_2}}$$

Where, s_1 = S.D. of the Ist sample, s_2 = S.D. of the 2nd sample, $SE_{s_1 - s_2}$ = Standard error of the difference between two S.D's.

Procedure : The following steps are taken for testing hypothesis about difference between two population standard directions :

(1) Set up the null hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.* there is no difference between the two population S.Ds.

Alternative hypothesis : $H_1 = \sigma_1 \neq \sigma_2$

(2) Compute the S.E.$s_1 - s_2$ by using the formulae.

(a) $SE_{s_1 - s_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$, when Population S.Ds of σ_1 and σ_2 are given.

(b) $SE_{s_1 - s_2} = \sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}$, when S.D s_1 and s_2 of the two samples are given.

(3) Substituting the values of s_1, s_2 and $SE_{s_1 - s_2}$ in the above stated Z-statistic.

The other stages such as (i) correct significance, (ii) critical value of Z, (iii) Decision including for testing hypothesis of the difference between two SD's are the same as given in the testing hypothesis about population mean μ.

Example 11. **The mean yield of two sets of plots and their variability are as given below, Examine whether the difference in the variability in yeelds is significant at 5% level of significance.**

	Set of 40 plots	Set of 60 plots
Mean yield per plot	1258 lbs	1243 lbs
S.D. per plot	34	28

Solution. We are given : $n_1 = 40, n_2 = 60, \bar{x}_1 = 1258$ lbs, $\bar{x}_2 = 1243$ lbs, $s_1 = 34$ lbs, $s_2 = 28$ lbs.

Null Hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.*, there is no significant difference in the variability in the yields between two sets at plots.

Alternative Hypothesis. $H_1 : \sigma_1 \neq \sigma_2$ (Two tailed test)

Level of significance $\alpha = 0.05$.

Test statistic : Under H_0 the test staristics, for large samples is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_1}}} \sim N(0, 1)$$

$$Z = \frac{34 - 28}{\sqrt{\frac{(34)^2}{80} + \frac{(28)^2}{120}}} = \frac{6}{\sqrt{\frac{1156}{80} + \frac{784}{120}}}$$

$$= \frac{6}{\sqrt{14.45 + 6.53}} = \frac{6}{\sqrt{20.98}} = \frac{6}{4.58} = 1.31$$

Since, $Z < 1.96$, it is not significant at 5% level of significance Hence, we may conclude that there is no siginficant difference in the variability in yields.

EXERCISE – 3

1. Random samples drawn from two countries gave the following data relating to the heights of adult males :

	Country A	Country B
Mean height in inches	67.42	67.25
Standard deviation in inches:	2.58	2.50
Number in samples	1000	1200

(i) Is the difference between the means significant ?

(ii) Is the difference between the standard deviations significant ?

[Ans. (i) $|Z| = 1.56$. Not significant. (ii) $|Z| = 1.03$. Not significant.]

2. The standard deviation of the height of B.A. (Hons.) students of a college is 4.0". Two samples are taken. The standard deviation of the height of 100 B. Com. Hons. students is

3.5" and B.A. Econ. Hons. students is 4.5". Test the significance of the difference of standard deviations of the samples.

[Ans. $H_0 : \sigma_1 = \sigma_2$; $H_1 : \sigma_1 \neq \sigma_2$; $|Z| = 2.89$; Significant.]

3. Intelligence test given to two groups of boys and girls gave the follwoing results :

Girls : Mean Marks = 78, S.D. = 12, $N = 80$

Boys : Mean Marks = 75, S.D. = 15, $N = 120$

(i) Is the difference is the mean scores significant?

(ii) Is the difference between standard deviation significant ?

[Ans. (i) Accept H_0, (ii) Reject H_0]

---

(4) Test of Hypothesis about Population Proportion : A random sample of size n ($n \geq 30$) has a sample proportion p of members possessing a certain attribute (*i.e.*, proportion of success). To test the hypothesis that the population proportion P has a specified value, the appropriate test statistic to be used as :

$$Z = \frac{p - P}{SE_p}$$

Where, p = Sample proportion, P = Population proportion.

SE_p = Standard error of proportion.

Procedure : The following steps are taken for test of hypothesis about population proportion :

(i) Set up the null hypothesis $H_0 : P = P_0$ *i.e.*, there is no difference between the sample proportion and population proportion.

Alternative hypothesis $H_1 : P \neq P_0$ ($\Rightarrow$ Two tailed test)

or $H_1 : P > P_0$ or $H_1 : P < P_0$ (One tailed test)

(ii) Compute the SE_p by using the following formuale :

(a) $SE_p = \sqrt{\frac{PQ}{n}}$ and where P is population proportion and n is the sample size.

(b) $SE_p = \sqrt{\frac{pq}{n}}$ and $q = 1 - p$ where P is not known.

(iii) Substituting the values of p, P and SE_p in the above stated Z-statistic.

The other steps such as (i) Level of significance, (ii) Critical value of Z_α, (iii) Decision making for testing hypothesis of the population proportion P are the same as those given in test of hypothesis about the population mean μ.

Example 12. **A coin is tossed 100 times under identical conditions independently yielding 30 heads and 70 tails. Test at 1% level of significance whether or not the coin is unbiased.**

Solution: **Here, the sample proportion p = Proportion to heads in the sample** $= \frac{30}{100} = 0{\cdot}30$

Also the population proportion $P = \frac{1}{2} = 0{\cdot}50 \Rightarrow Q = 1 - 0{\cdot}50 = 0{\cdot}50$

Null hypothesis $H_0 : P = 0{\cdot}5$ (That is, coin is unbiased)

Alternative hypothesis $H_1 : P \neq \frac{1}{2}$ ($\Rightarrow$ Two tailed test)

$$SE_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{{\cdot}50 \times {\cdot}50}{100}} = 0{\cdot}05$$

$$|Z| = \frac{|p - P|}{SE_p} = \frac{|0{\cdot}30 - {\cdot}50|}{{\cdot}05} = \frac{0{\cdot}20}{{\cdot}05} = 4$$

At 1% level of significance, the critical value of Z for two tailed test = 2·58.

Since, the calculated value of $|Z|$ > the critical value of Z, we reject the null hypothesis and hence conclude that the coin is biased.

Aliter : This question can also be solved on the basis of number of successes as follows :

Given : $n = 100$, $P = P(H) = \frac{1}{2} = 0{\cdot}5; Q = 1 - P = 1 - 0{\cdot}5 = 0{\cdot}5$

$H_0 = nP = 50$, $H_1 = nP \neq 50$

$\alpha = {\cdot}01$, $Z = 2{\cdot}58$ (Critical value), $np = 30$

$$\text{S.E.}_{np} = \sqrt{nPQ} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{25} = 5$$

$$|Z| = \frac{|np - nP|}{\text{S.E.}_{np}} = \frac{|30 - 50|}{5} = \frac{20}{5} = 4$$

Since, the calculated value of Z > the critical value of Z, we reject the null hypothesis and hence conclude that the coin is biased.

Example 13. **In 324 throws of six-faced dice, odd points appeared 180 times. Would you say that the die is fair ? Use $\alpha = 0{\cdot}05$.**

Solution. Here the sample proportion $p = \frac{180}{324} = 0{\cdot}555$

Also the population proportion $P = \frac{1}{2} = 0{\cdot}50 \Rightarrow Q = 1 - 0{\cdot}5 = 0{\cdot}50$

Null hypothesis $H_0 : P = 0{\cdot}5$

Alternative hypothesis $H_1 : P \neq 0{\cdot}5$ ($\Rightarrow$ Two tailed test)

$$\text{S.E.}_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{{\cdot}50 \times {\cdot}50}{324}} = \sqrt{\frac{0{\cdot}25}{324}} = 0{\cdot}027$$

Using Z-statistic, we have

$$|Z| = \frac{p - P}{\text{S.E.}_p} = \frac{0{\cdot}555 - {\cdot}50}{{\cdot}027} = \frac{{\cdot}055}{{\cdot}027} = 2{\cdot}03$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of Z > the critical value of Z, we reject the null hypothesis and conclude that the die is not fair.

Aliter : This question can also be solved on the basis of number of successes as follows :

Given : $n = 324$, $P = P$ (odd points) $= \frac{1}{2} = 0{\cdot}5; Q = 1 - P = 1 - 0{\cdot}5 = 0{\cdot}5$

$H_0 : nP = 162,$

$H_1 : nP \neq \frac{1}{2}, \alpha = {\cdot}05$

At $\alpha = {\cdot}05$, $Z = 1{\cdot}96$ (Critical value), $np = 180$

$$\text{S.E.}_{np} = \sqrt{nPQ} = \sqrt{324 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{81} = 9$$

$$|Z| = \frac{|np - nP|}{\text{S.E.}_{np}} = \frac{|180 - 162|}{9} = \frac{18}{9} = 2$$

Since, the conclulated value of Z > the critical value of Z, we reject the null hypothesis and hence conclude that the die is not fair.

Example 14. **In a sample of 500 persons from a village in Haryana, 280 are found to be rice eaters and the rest wheat eaters. Can we assume that both the food articles are equally popular ?**

Solution. Here the sample proportion $= p = \frac{280}{500} = 0{\cdot}56$

(*i.e.*, Proportion of rice eaters in the sample)

Also the population proportion $= P = \frac{1}{2} = 0{\cdot}50 \quad \Rightarrow \quad 1 - 0{\cdot}50 = 0{\cdot}50$

Null hypothesis $H_0 : P = \frac{1}{2}$ *i.e.*, both the food articles are equally popular

Alternative hypothesis $H_1 : P \neq \frac{1}{2}$ ($\Rightarrow$ Two tailed test)

$$\text{S.E.}_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{{\cdot}50 \times {\cdot}50}{500}} = 0{\cdot}022$$

Using Z-statistic, we have

$$|Z| = \frac{p - P}{\text{S.E.}_p} = \frac{0{\cdot}56 - 0{\cdot}50}{{\cdot}022} = \frac{0{\cdot}06}{{\cdot}022} = 2{\cdot}727$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since the concluded value of Z > the critical value of Z, we reject H_0 in favour of H_1 and hence both the food articles are not equally popular.

Aliter : This question can also be solved on the basis of number of successes as follows :

Given : $n = 500$, $P = P$ (rice eater) $= \frac{1}{2} = 0{\cdot}5; Q = 1 - P = 1 - 0{\cdot}5 = 0{\cdot}5$

$H_0 : nP = 250,$

$H_1 : nP \neq 250, \alpha = {\cdot}05$

At $\alpha = {\cdot}05$, $Z = 1{\cdot}96$ (Critical value), $np = 280$

$$S.E._{np} = \sqrt{nPQ} = \sqrt{500 \times \frac{1}{2} \times \frac{1}{2}} = 11{\cdot}18$$

$$|Z| = \frac{|np - nP|}{S.E._{np}} = \frac{|280 - 250|}{11{\cdot}18} = 2{\cdot}68$$

Since, the calculated value of Z > the critical value of Z, we reject the null hypothesis and hence conclude that both the food articles are not equally popular.

Example 15. **In a hospital, 480 female and 520 male babies were born in a week. Do these figures confirm the hypothesis that males and females are born in equal number ?**

Solution. Here, the sample proportion $= p = \frac{480}{1000} = 0{\cdot}48$

Also, the population proportion $= P = \frac{1}{2} = 0{\cdot}50$

Null hypothesis $H_0 : P = \frac{1}{2}$ *i.e.*, male and female are born in equal number.

Alternative hypothesis $H_1 : P \neq \frac{1}{2}$ ($\Rightarrow$ Two tailed test)

$$S.E._p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{\cdot}50 \times {\cdot}50}{1000}} = 0{\cdot}0158$$

Using Z-statistic, we have

$$|Z| = \frac{|p - P|}{S.E._p} = \frac{|0{\cdot}48 - 0{\cdot}50|}{0{\cdot}0158} = \frac{{\cdot}02}{{\cdot}0158} = 1{\cdot}26$$

At 5% level of significance, the critical value of Z for two tailed test $= 1{\cdot}96$.
Since, the calculated value of $Z <$ critical value of Z, we accept H_0. This means that the figures support the hypothesis that males and females are born in equal number.

Aliter : This question can also be solved on the basis of number of successes as follows :

Given : $n = 1000$, $P = P$ (Female) $= \frac{1}{2} = 0{\cdot}5$; $Q = 1 - P = 1 - 0{\cdot}5 = 0{\cdot}5$

$H_0 : nP = 500$, $H_1 : nP \neq 500$, $\alpha = {\cdot}05$

At $\alpha = {\cdot}05$, $Z = 1{\cdot}96$ (Critical value), $np = 480$

$$S.E._{np} = \sqrt{nPQ} = \sqrt{1000 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{250} = 15{\cdot}81$$

$$|Z| = \frac{|np - nP|}{S.E._{np}} = \frac{|480 - 500|}{15{\cdot}81} = 1{\cdot}26$$

Since, the calculated value of $Z <$ the critical value of Z, we accept the null hypothesis and hence conclude that males and females are born in equal numbers.

Example 16. ***A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler at 5% level of significance.***

Solution. Here, the sample proportion = p = proportion of defective apples = $\frac{36}{600} = 0{\cdot}06$.

Also, the population proportion = $P = 4\% = {\cdot}04 \Rightarrow Q = 1 - {\cdot}04 = {\cdot}96$

Null hypothesis $H_0 : P = 4\%$ or ·04

Alternative hypothesis $H_1 : P \neq 4\%$ or ·04

It is a case of two tailed test

$$\text{S.E.}_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{\cdot}04 \times 0{\cdot}96}{600}} = 0{\cdot}008$$

Using Z-statistic, we have

$$|Z| = \frac{|p - P|}{\text{S.E.}_p} = \frac{|0{\cdot}06 - 0{\cdot}04|}{{\cdot}008} = \frac{0{\cdot}02}{{\cdot}008} = 2{\cdot}5$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of Z > the critical value of Z, we reject H_0 in favour of H_1. It means that the claim of the wholesaler cannot be accepted.

Example 17. **A manufacturer claimed that 95% of the equipment which are supplied to a factory conform to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at a level of significance (a) ·05 and (b) 0·01.**

Solution. Here, the sample proportion

p = Proportions of pieces conforming to specifications para

$$= \frac{200 - 18}{200} = \frac{182}{200} = 0{\cdot}91.$$

Also population proportion = $P = 95\%$ or $0{\cdot}95 \Rightarrow Q = 1 - 0{\cdot}95 = 0{\cdot}05$

Null hypothesis $H_0 : P \geq 0{\cdot}95$ (Claim is justified)

Alternative hypothesis $H_1 : P < 0{\cdot}95$ (Claim is not justified)

It is a case of left tailed test.

$$\text{S.E.}_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{\cdot}95 \times 0{\cdot}05}{200}} = 0{\cdot}0154$$

Using Z-statistic, we have

$$|Z| = \frac{|p - P|}{\text{S.E.}_p} = \frac{|0{\cdot}91 - 0{\cdot}95|}{{\cdot}0154} = 2{\cdot}6$$

At 5% level, the critical value of Z for left tailed test = 1·645

At 1% level, the critical value of Z for left tailed test = 2·33

Since, the calculated value of Z > the critical value of Z for one tailed test both at 5% and 1%, we reject H_0 in favour of H_1. It means that the manufacture's claim is rejected.

Example 18. **In a big city, 450 men out of a sample of 850 were found to be smokers. Does this information support the view that the majority of men in the city are smokers ? Assume $\alpha = {\cdot}01$.**

Solution. Here, sample proportion = p = proportion of smokers = $\frac{450}{850} = 0{\cdot}53$

And, population proportion = $P = \frac{1}{2} = 0{\cdot}50$

Also $Q = 1 - P = 1 - 0{\cdot}5 = 0{\cdot}5$

Null hypothesis $H_0 : P = 0{\cdot}50$ (That is, smokers and non-smokers are equal in numbers)

Alternative hypothesis $H_1 : P > 0{\cdot}50$ ($\Rightarrow$ Right tailed test)

$$\text{S.E.}_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{\cdot}50 \times 0{\cdot}50}{850}} = 0{\cdot}0171$$

Using Z-statistic, we have

$$|Z| = \frac{p - P}{\text{S.E.}_p} = \frac{0{\cdot}53 - 0{\cdot}50}{{\cdot}00171} = \frac{{\cdot}03}{0{\cdot}0171} = 1{\cdot}754$$

At 1% level of significance, the critical value of Z for right tailed test = 2·33.

Since, the calculated value of $|Z|$ < the critical value of Z, we accept H_0 and conclude that majority of mean in city are not smokers.

Example 19. **A manufacturer claims that a shipment of finished nails contains less than 2% defective nails. A random sample of 400 nails when examined for defective items, is found to be containing 16 defectives. Test the claim at $\alpha = 0{\cdot}05$ level of significance.**

Solution. Here, the sample proportion = p = proportion of defective nail = $\frac{16}{400} = 0{\cdot}04$

And, population proportion = P = 2% or $0{\cdot}02 \Rightarrow Q = 1 - P = 1 - 0{\cdot}02 = 0{\cdot}98$

Null hypothesis $H_0 : P < 0{\cdot}02$ *i.e.*, his claim is accepted.

Alternative hypothesis $H_1 : P \geq 0{\cdot}02$ ($\Rightarrow$ right tailed test)

$$\text{S.E.}_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{\cdot}02 \times 0{\cdot}98}{400}} = 0{\cdot}007$$

Using Z-statistic, we have

$$|Z| = \frac{p - P}{\text{S.E.}_p} = \frac{0{\cdot}04 - 0{\cdot}02}{0{\cdot}007} = 2{\cdot}857$$

At 5% level of significance, the critical value of Z for one tailed test = 1·645.

Since, the calculated value of Z > the critical value of Z for one tailed test, we reject the null hypothesis and conclude that the manufacturer's claim is not acceptable.

EXERCISE – 4

1. *A coin is tossed 10,000 times and head turns up 5195 times. Would you consider the coin unbiased ?* [**Ans.** $Z = 3{\cdot}9$, H_0 is rejected]
2. *A die is thrown 49152 times and out of these 25145 yielded either 4 or 5 or 6. Is this consistent with the hypothesis that the die is unbiased.* [**Ans.** $Z = 5{\cdot}37$, H_0 is rejected]

3. A die was thrown 9000 times and out of these 3220 yielded a 3 or 4. Is this consistent with the hypothesis that the die was unbiased ? [**Ans.** $Z = 5$, H_0 is
4. A sales clerk in the departmental store claims that 60% of the shoppers entering the store leave without making a purchase. A random sample of 50 shoppers showed that 35 of them left without buying any thing. Are these sample results consistent with the claim of the sale clerk ? Use a 5% level of significance. [**Ans.** $Z = 1{\cdot}44$, H_0 is accepted]
5. In a sample of 400 parts manufactured by a company, the number of defective parts were found to be 30. The company, however claimed that only 5% of their product is defective. Test at 5% level of significance whether the claim of the company is tenable.

   [**Ans.** $Z = 2{\cdot}29$, reject H_0 *i.e.*, claim is not tenable]
6. A manufacturer claims that at least 90% of his goods supplied conform to specifications. A sample of 100 pieces has shown that 20 were faulty. Test his claim at 5% level of significance. [**Ans.** $|Z| = 3{\cdot}33$, H_0 is rejected]
7. In a big city, 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers [State the hypothesis clearly]. [**Ans.** $|Z| = 2{\cdot}04$, H_0 is rejected]
8. A medicine is claimed to be successful in 90% cases. In a sample of 200 patients 160 were cured. Will you accept the claim at 1% level of significance ? [**Ans.** $Z = 4{\cdot}71$, reject H_0]

---

(5) Test of Hypothesis about the difference between two population Proportions : Let p_1 and p_2 be the sample proportions obtained in large sample of sizes n_1 and n_2 drawn from respective populations having proportions P_1 and P_2. To test the hypothesis that there is no difference between the two population proportions, the appropriate test statistic to be used is :

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\text{S.E}_{p_1 - p_2}}$$

Where, p_1 = 1st sample proportion, p_2 = 2nd sample proportion

$\text{S.E}_{p_1 - p_2}$ = Standard error of the difference of two proportions.

Procedure : The following steps are taken for test of hypothesis about the difference between two population proportions :

(i) Set up the null hypothesis $H_0 : P_1 = P_2$ *i.e.*, there is no difference betweeen two population proportions.

Alternative hypothesis : $H_1 : P_1 \neq P_2$ (Two tailed test)

or $H_1 : P_1 > P_2$ or $P_1 < P_2$ (One tailed test)

(ii) Compute $\text{S.E}_{p_1 - p_2}$ by using any one of the following :

(a) When the population proportions P_1 and P_2 are known, then

$$\text{S.E}_{p_1 - p_2} = \sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}$$

Where, n_1 and n_2 are the sizes of the two sampes.

(b) When the population proportion P_1 and P_2 are not known but sample proportions p_1 and p_2 are known.

$$\text{S.E}_{p_1 - p_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \text{Where } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} \text{ so that } q = 1 - p$$

(P = pooled estimate)

(iii) Substituting the values of p_1, p_2, P_1, P_2 and $S.E._{p_1-p_2}$ in the above stated Z-statistic.

The other steps such as (i) level of significance, (ii) critical value of Z_α, (iii) Decision making for testing the hypothesis of the difference between two proportions are the same as those given in test of hypothesis.

Example 20. **In a certain district A, 450 persons were considered regular consumers of tea out of a sample of 1000 persons. In another district B, 400 persons were regular consumers of tea out of a sample of 800 persons. Do these data indicate a significant difference between the two districts so far as drinking habit is concerned ? (Use 5% level)**

Solution. Let P_1 and P_2 be the population proportions of persons who are regular consumers of tea in the districts A and B respectively.

We set up null hypothesis $H_0 : P_1 = P_2$ *i.e.*, there is no significant difference in tea drinking habits in two districts.

Alternative hypothesis $H_1 : P_1 \neq P_2$ ($\Rightarrow$ Two tailed test)

$$n_1 = 1000, p_1 = \frac{450}{1000} = 0{\cdot}45, n_2 = 800, p_2 = \frac{400}{800} = 0{\cdot}5$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{450 + 400}{800 + 1000} = \frac{850}{1800} = 0{\cdot}47$$

and $q = 1 - p = 1 - 0{\cdot}47 = 0{\cdot}53$

$$S.E._{p_1-p_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0{\cdot}47)\,(0{\cdot}53)\left(\frac{1}{1000} + \frac{1}{800}\right)}$$

$$= \sqrt{0{\cdot}00056} = 0{\cdot}0237$$

Now, using Z-statistic we have

$$|Z| = \frac{p_1 - p_2}{S.E._{p_1-p_2}}$$

$$= \frac{0{\cdot}45 - 0{\cdot}50}{0{\cdot}0237} = \frac{0{\cdot}05}{0{\cdot}0237} = 2{\cdot}1097$$

At 5% level, the critical value Z for two tailed test = 1·96.

Since, the calculated value of Z > the critical value of Z, we reject the H_0 and conclude that there is significant difference between the two districts so far as tea-drinking habit is concerned.

Example 21. **A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved ?**

Solution. Let P_1 and P_2 be the proportions of defective articles in the population of articles manufactured by the machine before the after overhauling, respectively.

We set up the null hypothesis $H_0 : P_1 \leq P_2$.

Alternative hypothesis $H_1 : P_2 < P_1$ or $P_1 > P_2$ ($\Rightarrow$ One tailed test)

We have, $n_1 = 200, p_1 = \frac{20}{400} = 0{\cdot}05, n_2 = 300, p_2 = \frac{10}{300} = {\cdot}0333$

$$p=\frac{n_1p_1+n_2p_2}{n_1+n_2}=\frac{x_1+x_2}{n_1+n_2}=\frac{20+10}{400+300}=\frac{3}{70}=\frac{3}{70}, q=1-p=\frac{67}{70}$$

$$\text{S.E.}_{p_1-p_2}=\sqrt{\frac{3}{70}\times\frac{67}{70}\cdot\left(\frac{1}{400}+\frac{1}{300}\right)}=\sqrt{\frac{201}{4900}\times\frac{7}{1200}}$$

$$=\sqrt{\frac{201}{840000}}=\sqrt{0\cdot0002392}=0\cdot0155$$

Now, using Z-statistic, we have

$$|Z|=\frac{p_1-p_2}{\text{S.E.}_{P_1-P_2}}=\frac{0\cdot05-0\cdot033}{0\cdot0155}=\frac{0\cdot017}{0\cdot0155}=1\cdot096$$

At 5% level of significance, the critical value of Z for one tailed test = 1·645.
Since the calculated value of $|Z| = 1\cdot096 <$ the critical value of Z, we accept H_0 at 5% level of significance and conclude that machine has not improved.

Example 22. **Before an increase in excise duty on tea, 400 persons out of a sample of 500 persons were found to be tea-drinkers. After an increase in excise duty, 400 persons were known to be tea drinkers in a sample of 600 persons. Do you think that there has been a significant decrese in the consumption of tea after the increase in excise duty ?**

Solution. Let P_1 and P_2 be the proportions of tea drinkers in the population of persons before and after the increase in excise duty.

We set up the null hypothesis $H_0 : P_1 = P_2$.

Alternative hypothesis $H_1 : P_2 < P_1$ or $P_1 > P_2$ (⇒ One tailed test)

Now, $n_1 = 500, p_1 = \frac{400}{500} = 0\cdot8, n_2 = 600, p_2 = \frac{400}{600} = 0\cdot667$

$$p=\frac{n_1p_1+n_2p_2}{n_1+n_2}=\frac{x_1+x_2}{n_1+n_2}=\frac{400+400}{500+600}=\frac{800}{1100}=\frac{8}{11}$$

And $q=1-p=1-\frac{8}{11}=\frac{3}{11}$

$$\text{S.E.}_{P_1-P_2}=\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\sqrt{\frac{8}{11}\times\frac{3}{11}\left(\frac{1}{500}+\frac{1}{600}\right)}=0\cdot027$$

Using Z-statistic, we have

$$|Z|=\frac{p_1-p_2}{\text{S.E.}_{P_1-P_2}}=\frac{0\cdot8-0\cdot667}{0\cdot027}=4\cdot93$$

At 5% level of significance, the critical value of Z for one tailed test = 1·645.
Since, the calculated value of $|Z| = 4\cdot93 >$ critical value of Z = 1·645, we reject the null hypothesis and conclude that there is a significant decrease in the consumption of tea after the increase in excise only.

Example 23. **A sample survey of tax payers belonging to Business class and Professional class yielded the following results :**

	Business Class	Professional Class
Sample Size	$n_1 = 400$	$n_2 = 420$
Defaulters in tax payment	$x_1 = 80$	$x_2 = 65$

Test the hypothesis that the defaulters rate is the same for the two classes of tax payers. (Use $\alpha = \cdot 05$ level of significance).

Solution. Let P_1 and P_2 be the proportions of tax-defaulters in business and professional classes, respectively.

We set up the null hypothesis $H_0 : P_1 = P_2$ *i.e.*, there is no significant difference in the defaulters rate for two classes of tax payers.

Alternative hypothesis $H_1 : P_1 \neq P_2$ ($\Rightarrow$ Two tailed test)

We have $n_1 = 400, p_1 = \frac{80}{400} = 0{\cdot}20, n_2 = 420, p_2 = \frac{64}{420} = 0{\cdot}15$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{80 + 65}{400 + 420} = \frac{145}{820} = 0{\cdot}177$$

and $q = 1 - p = 1 - 0{\cdot}177 = 0{\cdot}823$

$$\text{S.E}_{P_1 - P_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0{\cdot}177 \times 0{\cdot}823\left(\frac{1}{400} + \frac{1}{420}\right)} = 0{\cdot}0266$$

Using Z-statistic, we have

$$|Z| = \frac{p_1 - p_2}{\text{S.E}_{P_1 - P_2}} = \frac{0{\cdot}20 - 0{\cdot}15}{0{\cdot}0266} = \frac{\cdot 05}{\cdot 0266} = 1{\cdot}87$$

At 5% level, the critical value of Z for two tailed test = 1·96.

Since, the calculated value of Z < the critical value of Z, we accept H_0 and conclude that there is no significant difference in the defaulters rate for two classes of tax payers.

Example 24. **A market researcher engaged by a particular company believes that the proportion of households using company's products in city A exceeds this proportion in city B by 0·05. The reseacher conducts survey of two cities and finds the following results :**

City A	Sample Size	No. of households using company's products
A	$n_1 = 160$	120
B	$n_2 = 150$	100

Use 0·05 level of significance and test the researcher's claim.

Solution. Let P_1 and P_2 the proportions of households using company's products in city A and city B.

We set up the null hypothesis $H_0 : P_1 - P_2 = 0{\cdot}05$

Alternative hypothesis : $P_1 - P_2 > 0{\cdot}05$ ($\Rightarrow$ One tailed test)

Now, $n_1 = 160, p_1 = \frac{120}{160} = 0{\cdot}75, n_2 = 150, p_2 = \frac{100}{150} = 0{\cdot}67$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 100}{160 + 150} = \frac{220}{310} = 0{\cdot}71$$

and $q = 1 - p = 1 - 0{\cdot}71 = 0{\cdot}29$

$$S.E._{P_1-P_2} = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0{\cdot}71\times 0{\cdot}29\left(\frac{1}{160}+\frac{1}{150}\right)} = 0{\cdot}0515$$

Now, using Z-statistic, we have

$$|Z| = \frac{(p_1-p_2)-(P_1-P_2)}{S.E._{P_1-P_2}} = \frac{(0{\cdot}75-0{\cdot}67)-(\cdot 05)}{0{\cdot}0515} = \frac{\cdot 03}{0{\cdot}0515} = 0{\cdot}582$$

At 5% level, the critical value of Z for one tailed test = 1·645.

Since, $|Z| = 0{\cdot}58 < 1{\cdot}645$, we accept H_0 and conclude that the proportion of households using the company's products in city A exceeds to that of city B by 0·05.

EXERCISE – 5

1. In a sample of 600 students of a certain college, 400 are found to use dot pens. In another college from a sample of 900 students, 450 were found to use dot pens. Test whether the two colleges are significantly different with respect to the habit of using dot pens at 5% level of significance. [Ans. $|Z| = 6{\cdot}51$ H_0 is rejected]
2. A machine produced 20 defective articles in a batch of 500. After overhauling it produced 3 defective articles in a batch of 100. Has the machine improved ?

   [Ans. $|Z| = \cdot 476$, H_0 is accepted]
3. Before an increase in excise duty on tobacco, 400 people out of 500 were found to be smokers. After an increase in the excise duty, 400 persons out of 600 were known to be smokers. Do you think that there has been a significant decrease in the proportions?

   [Ans. $|Z| = 4{\cdot}81$, H_0 is rejected]
4. In a sample of 600 men from a city, 450 are found to be smokers. In a sample of 900 from another city, 450 are found to be smokers. Do the data indicate that two cities differs significantly in their smoking habits ? [Ans. $|Z| = 9{\cdot}7$, H_0 is rejected]
5. From a random sample of 200 students from the Kurukshetra University, 90 were found to be copying and from a similar sample of 160 students from the M.D.U. Rohtak, 64 were found to be copying. Do these figures suppose the hypothesis that there is no significant difference between the two universities so far as the proportion of coyping students is concerned ? [Ans. $Z = 0{\cdot}95$, H_0 is accepted]
6. A sample survey of tax payers belonging to business class and professional class yielded the following results :

	Business Class	Professional Class
Sample Size	$n_1 = 200$	$n_2 = 160$
Defaulters in tax payment	$x_1 = 90$	$x_2 = 64$

   Test that the defaulters rate is the same for the two classes of tax payers.

   [Ans. $|Z| = 0{\cdot}95$, H_0 is accepted]
7. 500 units from a factory are inspected and 12 are found to be defective, similarly, 800 units from another factory are inspected and 12 are found to be defective can it be concluded at 5% level of significance that productions in second factory is better than in first factory.

   [Ans. $Z = 1{\cdot}184$, H_0 is accepted]
8. A survey of television audience in a big city revealed that a particular hight programme was liked by 50 out of 200 males and 80 out of 250 females. Test the hypothesis at $\alpha = \cdot 05$ level

of significance whether that is a great difference of opinion about the proramme between the males and females. [Ans. $|Z| = 1 \cdot 627$, H_0 is accepted]

9. In a random sample of 2,100 persons from Panjab 1260 persons are found to be honest. In another random sample of 4000 persons from Haryana 2360 persons are found to be honest. Do the data indicate at 1 percent level of significance that (a) the two cities are different with respect to honesty ? (b) the persons in Panjab are more honest as compared to Haryana. [Ans. (a) $H_0 : P_1 = P_2$; $H_1 : P_1 \neq P_2$, $|Z| = 0 \cdot 75$, Accept H_0
(b) $H_0 : P_1 \leq P_2$; $H_1 : P_1 > P_2$, $|Z| = 0 \cdot 75$, Accept H_0]

10. An advertising agency wants to find out if there is any difference in the degree of loyalty for a given brand of cereal between men and women. A randon sample of 200 men and 200 women was taken and it was determined that 58% of women and 65% of men showed brand loyalty. At $\alpha = 0.05$, test the null hypothesis that there is no significant difference between the population of men and women who are brand loyal.

[Ans. 2 = 1.47, Accept H_0]

11. Company is considering two different TV advertisements for promotion of a new product. Management believes that advertisement A is more effective than advertisement B. Two test market X and Y with virtually identical consumer characteristics were selected to test this belief. A is used in X and B is used in Y. In market X, out of 60 customers who saw the advertisement, 18 bought the product, while for market B, the respective figures were 10 and 22. Do the results indicate that advartisement A is more effective than advertisement B? Test at 5% level of significance.

(**Hint :** Use one tailed test)

[Ans. $Z = 1.133$, Accept H_0]

MISCELLANEOUS SOLVED EXAMPLES

Example 25. **A man buys 50 electric bulbs of each of 'Philips' and 'HMT'. He finds that Philips bulbs has a mean life of 1500 hours and a S.D. of 60 hours and HMT bulbs has a mean life of 1512 hours with a S.D. of 80 hours. Is there a significant difference in the mean life of these two kinds of bulbs ?**

Solution. Let us take the hypothesis that there is no significant difference in mean life of two makes of bulbs *i.e.*,

Null hypothesis $H_0 : \mu_1 = \mu_2$ And Altenative hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

Given : $n_1 = 500, \bar{X}_1 = 1500, s_1 = 60$

$n_2 = 50, \bar{X}_2 = 1512, s_2 = 80$

$$S.E_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(60)^2}{50} + \frac{(80)^2}{50}}$$

$$= \sqrt{\frac{3600}{50} + \frac{6400}{50}} = \sqrt{72 + 128}$$

$$= \sqrt{200} = 14 \cdot 14$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{S.E_{\bar{X}_1 - \bar{X}_2}} = \frac{|1500 - 1512|}{14 \cdot 14} = 0 \cdot 848$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.
Since. the calculated value of Z is less than 1·96 at 5% level of significance, we accept the null hypothesis and conclude that the difference in mean life of two kinds of bulbs is not significant.

Example 26. **A sample of 400 persons was taken from a large population. The mean weight and standard deviation of these persons was found to be 68 kg. and 6 kg. Can it reasonably be said that in the population mean weight will be 67 kg ?**

Solution. Let us take the hypothesis that the population mean is equal to 67 *i.e.*,
$H_0 : \mu = 67$ and $H_1 : \mu \neq 67$
Given : $n = 400, \bar{X} = 68, \mu = 67, s = 6$

$$\text{S.E.}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{6}{\sqrt{400}} = \frac{6}{20} = 0 \cdot 3$$

$$|Z| = \frac{\bar{X} - \mu}{\text{S.E.}_{\bar{X}}} = \frac{68 - 67}{0 \cdot 3} = \frac{1}{\cdot 3} = 3 \cdot 33$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.
Since, the calculated value of Z is more than 1·96 at 5% level of significanc, we reject the null hypothesis and conclude that population mean is not equal to 67.

Example 27. **The means of two large samples of size 400 and 900 are 75 and 75·75 respectively. Test the equality of the means of two populations each with S.D. of 2.**

Solution. Let us take the null hypothesis that there is no difference between the two population means *i.e.*,
$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu \neq \mu_2$ (⇒ Two tailed test)
Given : $n_1 = 400, n_2 = 900, \bar{X}_1 = 75, \bar{X}_2 = 75 \cdot 75, \sigma = 2$

$$\text{S.E.}_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = 2 \sqrt{\frac{1}{400} + \frac{1}{900}}$$

$$= 2 \cdot \sqrt{\frac{9 + 4}{3600}} = 2 \cdot \sqrt{\frac{13}{3600}} = 0 \cdot 12$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{S.E.}_{\bar{X}_1 - \bar{X}_2}} = \frac{|75 - 75 \cdot 75|}{0 \cdot 12} = 6 \cdot 25$$

At 5% level significance, the critical value of Z for two tailed test = 1·96.
Since, the calculated value of $|Z|$ is greater than the critical value of Z, we reject the hypothesis and conclude that the means differ significantly.

Example 28. **Intelligence test given to two groups of boys and girls gave the following results :**

	Mean Marks	S.D.	Number of Students
Girls	78	12	80
Boys	75	14	120

Is the difference in the mean scores significant ?

Solution. We are given :

$n_1 = 80, \quad \bar{X}_1 = 78, \quad s_1 = 12$

$n_2 = 120, \quad \bar{X}_2 = 75, \quad s_2 = 14$

Let us take the null hypothesis that the mean of two populations are equal *i.e.*,

$H_0 : \mu_1 = \mu_2$

Alternative hypothesis : $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$S.E._{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(12)^2}{80} + \frac{(14)^2}{120}}$$

$$= \sqrt{1 \cdot 8 + 1 \cdot 63} = 1 \cdot 852$$

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{S.E._{\bar{X}_1 - \bar{X}_2}} = \frac{78 - 75}{1 \cdot 852} = 1 \cdot 62$$

At 5% level of significance, the critical value of Z for two tailed test = 1·96.

Since, the calculated vlaue of $|Z| = 1 \cdot 62 <$ critical value of Z, we accept the null hypothesis and conclude that there is no significant difference between the means score of boys and girls.

Example 29. **An automobile manufacturer asserts that the seat belts of his firms are 90 percent effective. A consumer group tests the seat belt on 50 cars and find it effective on 37 of item. Test the correctness of the manufacturer assertion at 5% level of significance.**

Solution. Here, the sample proportion $= p = \frac{37}{50} = 0 \cdot 74$

Also, the population proportion $= P = 90\%$ or $0 \cdot 90 \Rightarrow Q = 1 - \cdot 90 = 0 \cdot 10$

Null hypothesis $H_0 : P = 90\%$ or 0·90 (Claim is justified)

Alternative hypothesis $H_1 : P < 0 \cdot 90$ (Claim is not justified)

It is a case of left tailed test

$$S.E._p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0 \cdot 90 \times \cdot 10}{50}} = 0 \cdot 04$$

$$|Z| = \frac{|p - P|}{S.E._p} = \frac{|0 \cdot 74 - 0 \cdot 90|}{\cdot 04} = 4$$

At 5% level of significance, the critical value of Z for left tailed test = 1·645.

Since, the calculated value of $|Z| >$ the critical value of Z, we reject H_0 and conclude that the automobile manufacture's assertion is not correct.

Example 30. **In two large populations, there are 30% and 25% respectively of curly haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations ?**

Solution. Let P_1 and P_2 be proportions of curly hair people in two large populations.

Given : $P_1 = 30\% = 0 \cdot 30, Q = 1 - 0 \cdot 30 = 0 \cdot 70, n_1 = 1200$

$P_2 = 25\% = 0 \cdot 25, Q_2 = 1 - 0 \cdot 25 = 0 \cdot 75, n_2 = 900$

We set up the null hypothesis $H_0 : P_1 = P_2$ *i.e.*, the difference in population likely to be hidden in the samples.

Alternative hypothesis $H_1 : P_1 \neq P_2$ ($\Rightarrow$ Two tailed test)

$$S.E._{P_1 - P_2} = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

[Here, proportion of successes in two populations are known]

$$= \sqrt{\frac{\cdot 3 \times \cdot 7}{1200} + \frac{\cdot 25 \times \cdot 75}{900}}$$

$$= \sqrt{\cdot 000175 + 0 \cdot 00021}$$

$$= \sqrt{\cdot 00385} = \cdot 0196$$

$$|z| = \frac{|P_1 - P_2|}{S.E._{P_1 - P_2}} = \frac{0 \cdot 30 - 0 \cdot 25}{\cdot 0196} = \frac{\cdot 05}{\cdot 0196} = 2 \cdot 55$$

At 5% level of significance, the critical value of Z for two tailed test $= 1 \cdot 96$.
Since, $|Z| = 2 \cdot 55 > 1 \cdot 96$, we reject H_0 and conclude that the difference is unlikely to be hidden due to simple sampling fluctuations.

Example 31. **The results of IQ test of two groups of girls are as under :**

Group of 120 Girls : Mean = 84, S.D. = 10

Group of 80 Girls : Mean = 81, S.D. = 12

Do they belong to the same population?

Solution. To test if two independent samples belong to the same population, we have to test :

(i) The equality of population means and
(ii) The equality of population standard deviations

(i) Let us take the null hypothesis that the mean of two groups of girls is not significant *i.e.* $H_0 : \mu_1 = \mu_2$

$x_1 = 120; n_2 = 80, s_1 = 10, s_2 = 12, \bar{X}_1 = 84, \bar{X}_2 = 81$

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(10)^2}{120} + \frac{(12)^2}{80}} = \sqrt{\frac{100}{120} + \frac{144}{80}}$$

$$= \sqrt{0.833 + 1.80} = \sqrt{2.633} = 1.62$$

Applying Z-statistic

$$|Z| = \frac{|\bar{X}_1 - \bar{X}_2|}{SE_{\bar{X}_1 - \bar{X}_2}} = \frac{|84 - 81|}{1.62} = \frac{3}{1.62} = 1.85$$

At 5% level of significance, the critical value of Z for two tailed test = 1.96
Since, the calculated value of $|Z|$ is less than the critical value of Z of 5% l.o.s, we accept the null hypothesis and conclude that there is no significant difference in the mean of two groups of girls.

(ii) Let us take the hypothesis that there is no difference in the S.Ds of the two groups of girls, *i.e.*

$H_0' : \sigma_1 = \sigma_2$

$$SE_{\sigma_1-\sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}} = \sqrt{\frac{(10)^2}{2\times 120}+\frac{(12)^2}{2\times 80}}$$

$$= \sqrt{\frac{100}{240}+\frac{144}{160}} = \sqrt{0.416+0.90}$$

$$= \sqrt{1.316} = 1.147$$

$$|Z| = \frac{|\sigma_1-\sigma_2|}{SE_{\sigma_1-\sigma_2}} = \frac{|10-12|}{1.147} = 1.74$$

Since, the calculated value of $|Z|$ is less than the critical value of Z at 5% l.o.s, we accept H_0 and conclude that there is no significant difference in the standard deviations of two groups of girls.

Since, both the Hypothesis $H_0 : \mu_1 = \mu_2$ and $H_0' : \delta_1 = \delta_2$ are accepted, we may say that the two groups of girls belong to the same population.

IMPORTANT FORMULAE

(i) Test of Hypothesis about population mean :

$$Z = \frac{\bar{X}-\mu}{S.E._{\bar{X}}}$$

Where $S.E._{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$

(ii) Test of Hypothesis about the difference between two population mean :

$$|Z| = \frac{\bar{X}_1-\bar{X}_2}{S.E._{\bar{X}_1-\bar{X}_2}}$$

Where $S.E._{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ ——

$$= \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$= \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

(iii) Test of Hypothesis about population proportion :

$$|Z| = \frac{p-P}{S.E._p}$$

Where $S.E._p = \sqrt{\frac{PQ}{n}}$

(iv) Test of Hypothesis about difference between two population proportions :

$$|Z| = \frac{p_1-p_2}{S.E._{p_1-p_2}}$$

Where $S.E._{P_1 - P_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ and $q = (1 - p)$

(p = polled estimate)

QUESTIONS

1. What is statistical hypothesis ? Discuss the procedure of testing a statistical hypothesis.
2. Explain the following :
   (a) Null hypothesis and alternative hypothesis.
   (b) Type I and Type II Errors.
   (c) One tail and two tailed test.
   (d) Acceptance and rejection regions
   (e) Level of significance.
3. How do you test the equality of two population means and two population proportions in case of large samples ?
4. Explain how the size of α (alpha) is determined for testing an hypothesis.
5. Outline the procedure for large sample tests and discuss their theoretical basis.
6. Discuss the applications of large sample tests.
7. Distinguish between a null hypothesis and an alternative hypothesis. Use example to explain the nature of null and alternative hypothesis in cases of one and two tailed tests.
8. Explain the concept of level of significance in test of hypothesis. Discuss briefly the procedure of testing a hypothesis.
9. State the null and alternative hypothesis regarding the population mean that lead to :
   (i) Left-tailed test (ii) Right-tailed test (iii) Two tailed test;

   [**Hint :** Left tailed test : $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$, Right-tailed : $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$; Two tailed test $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$].

❀❀❀

4

Tests of Hypothesis – Small Sample Tests

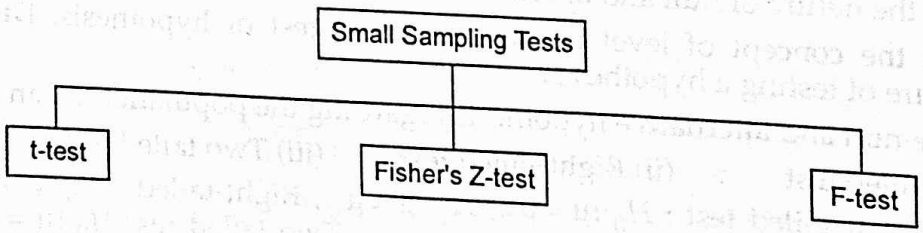
INTRODUCTION

The various tests of hypothesis or test of significance discussed in the previous chapter were related to large samples. Tests of hypothesis relating to large samples are based on two assumptions : (a) Sampling distribution of a statistic approaches a normal distribution whether or not the population distribution is normal or not, (b) The values of the sample statistics are sufficient close to the population values. But these two basic assumptions of large samples don't hold good in case of small samples. Therefore, it becomes necessary to make a separate study of small sample tests. Various small sampling tests are available; the most common being them are :

(1) *t*—test

(2) Fisher's Z—test

(3) *F*—test



(1) *t*—test

t—test is a small sample test. It was developed by William Gosset in 1908. He published this test under the pen name of "Student". Therefore, it is known as Student's *t*–test. For applying *t*–test, the value of *t*–statistic is computed. For this, the following formula is used :

$$t = \frac{\text{Deviation from the population parameter}}{\text{Standard Error of the sample statistic}}$$

The calculated value of *t* is compared with the table value of *t* for given degrees of freedom at certain specified level of significance.

Applications/Uses of *t*—test : The following are some important applications of *t*–test :

(1) Test of hypothesis about the population

(2) Test of hypothesis about the difference between the two means in case of independent samples.

(3) Test of hypothesis about the difference between two means with dependent samples.

(4) Test of hypothesis about an observed coefficient of correlation.

(1) Test of hypothesis about the population mean (σ unknown and sample size is small) : A random sample of size n (≤ 30) drawn from a normal population has a sample mean $\bar{X}$. To test the hypothesis that the population mean μ has a specified value μ_0 when population standard deviation σ is not known, and $n < 30$, we use t-test and the appropriate test statistic t to be used is :

$$t = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

where, S = Modified S.D. of the sample.
n = Size of the sample

Procedure : The following steps are taken while testing the hypothesis about the population mean μ :

(1) Set up the null hypothesis $H_0 : \mu = \mu_0$ *i.e.*, the population mean is μ_0
Alternative hypothesis $H_1 : \mu \neq \mu_0$ ($\Rightarrow$ Two tailed test)
or $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ ($\Rightarrow$ One tailed test)

(2) Thereafter the value of modified standard deviation of the sample (S) is computed by using any of the following formulae :
(a) When deviations are taken from the actual mean :

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

(b) When actual mean is in fraction and deviations are taken from assumed mean :

$$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$$

Where, $d = X - A$ (deviation from the Assumed Mean)
(c) When sample standard deviation is given :

$$S = \sqrt{\frac{n}{n-1}} s^2 \quad \text{or} \quad \sqrt{\frac{n}{n-1}} \cdot s$$

(3) The values of $\bar{X}$, μ and S are substituted in the above stated formula.
(4) Degrees of freedom are worked out by using the following formula :
Degrees of Freedom $= v = n - 1$
(5) Obtain the table value of t at the stated level of significance (α) and for given degrees of freedom from the table of "values t-statistics"
(6) If the computed value of $t = |t| <$ table value of t at a level of significance α, then we accept the null hypothesis.

If the computed value of $t >$ table value of t at a level of significance α, then we reject the null hypothesis and accept the alternative hypothesis.

Important Note : Since 't' distribution is symmetric about $t = 0$, the significant values at a level of significance 'α' for a single tailed (right or left) test can be obtained from the table of two tailed test by looking the value at the level of significance 2α. For example :

t^v (0·05) for single tailed test = t^v (0·10) for two tailed test.

t^v (0·01) for single tailed test = t^v (0·02) for two tailed test.

t^v (0·025) for single tailed test = t^v (0·05) for two tailed test.

t^v (0·005) for single tailed test = t^v (0·01) for two tailed test.

The testing procedure of population mean is clarified from the following examples :

Example 1. **A group of 5 patients treated with medicine *A* weights : 42, 39, 48, 60 and 41 kg. In the light of the above data, discuss the suggestion that mean weight of the population is 48 kg. Test at 5% level of significance.**
(Given the table value of *t* for 4 d.f. at 5% level is 2·776)

Solution. Let us take the null hypothesis that mean weight in the population is 48 *i.e.*,

$H_0 : \mu = 48$ and $H_1 : \mu \neq 48$ (⇒ Two tailed test)

Weight (X)	$\bar{X} = 46$ $(X - \bar{X})$	$(X - \bar{X})^2$
42	−4	16
39	−7	49
48	2	4
60	14	196
41	−5	25
$\Sigma X = 230, n = 5$		$\Sigma(X - \bar{X})^2 = 290$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{230}{5} = 46$$

$$S = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n-1}} = \sqrt{\frac{290}{5-1}} = \sqrt{\frac{290}{4}} = 8{\cdot}514$$

Applying *t*–test :

$$t = \frac{\bar{X} - \mu}{S} . \sqrt{n}$$

$$|t| = \frac{46 - 48}{8{\cdot}514} . \sqrt{5} = \frac{2 \times 2{\cdot}236}{8{\cdot}514} = \frac{4{\cdot}472}{8{\cdot}514} = 0{\cdot}525$$

Degrees of freedom $= v = n - 1 = 5 - 1 = 4$

For $v = 4$, $t_{0{\cdot}05}$ for two tailed test $= 2{\cdot}776$

Since, the calculated value of *t* is less than the table value, we accept the null hypothesis and therefore conclude that the mean weight in the population is 48 kg.

Example 2. **A random sample of 9 boys had heights (inches) : 45, 47, 50, 52, 48, 47, 49, 53 and 51. In the light of the data, discuss the suggestion that the mean height in the population is 47·5.**
(Give the table value of *t* for 8 d.f. at 5% level = 2·306).

Solution. Let us take the null hypothesis that the mean height in the population is 47·5 *i.e.*,

$H_0 : \mu = 47{\cdot}5$ and $H_1 = \mu \neq 47{\cdot}5$ (⇒ Two tailed test)

Height X	$A = 49$ $d = (X - A)$	d^2
45	−4	16
47	−2	4
50	+1	1
52	+3	9
48	−1	1
47	−2	4
49	0	0

1 tail Test = 1.92 @ 5%
2 tail Test = 8.776 @ 5%.

53 51	+4 +2	16 4
$\Sigma X = 442$	$\Sigma d = 1$	$\Sigma d^2 = 55$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{442}{9} = 49{\cdot}11$$

Since, the actual mean is in fraction, we should take deviations from assumed mean (49) to simplify the calculations.

$$\bar{d} = \frac{\Sigma d}{n} = \frac{1}{9} = 0{\cdot}11$$

$$S = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{55 - 9 \times (\cdot 11)^2}{9-1}} = \sqrt{\frac{54{\cdot}8911}{8}} = \sqrt{6{\cdot}8613} = 2{\cdot}62$$

Applying *t*–test :

$$t = \frac{\bar{X} - \mu}{S} . \sqrt{n} = \frac{49{\cdot}11 - 47{\cdot}5}{2{\cdot}62} \sqrt{9}$$

$$= \frac{1{\cdot}61 \times 3}{2{\cdot}62} = \frac{4{\cdot}83}{2{\cdot}62} = 1{\cdot}843$$

Degrees of freedom $= v = 9 - 1 = 8$

For $v = 8$, $t_{0{\cdot}05}$ for two tailed test $= 2{\cdot}306$

Since, the calculated value of t is less than the table value, we accept the null hypohtesis and therefore conclude that the mean height in the population is 47·5 inches.

Aliter : The vlaue of S can also be calculated as :

$$\Sigma (X - \bar{X})^2 = \Sigma d^2 - \frac{(\Sigma d^2)}{n} = 55 - \frac{(1)^2}{9} = 54{\cdot}89$$

$$\therefore \quad S = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n-1}} = \sqrt{\frac{54{\cdot}89}{9-1}}$$

$$= \sqrt{\frac{54{\cdot}89}{8}} = 2{\cdot}619 \simeq 2{\cdot}62$$

Example 3. **A random sample of 9 values from a normal population showed a mean of 41·5 inches and the sum of squares of deviations from this mean equal to 72 inches. Show whether the assumption of mean of 44·5 inches in the population is reasonable. (For $v = 8$, $t_{.05} = 2{\cdot}776$)**

Solution. $\bar{X} = 41{\cdot}5, \mu = 44{\cdot}5, n = 9, \Sigma (X - \bar{X})^2 = 72$

Let us take the null hypothesis that the population mean is 44·5

i.e., $H_0 : \mu = 44{\cdot}5$ and $H_1 : \mu \neq 445$

$$S = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n-1}} = \sqrt{\frac{72}{9-1}} = \sqrt{\frac{72}{8}} = \sqrt{9} = 3$$

Applying *t*–test :

$$|t| = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

$$|t| = \frac{|41{\cdot}5 - 44{\cdot}5|}{3} \times \sqrt{9} = 3$$

Degrees of freedom $= v = n - 1 = 9 - 1 = 8$

For $v = 8$, $t_{0{\cdot}05}$ for two tailed test $= 2{\cdot}306$

Since, the calculated value of $|t|$ > the table value of t, we reject the null hypothesis. We conclude that the population mean is not equal to 44·5.

Example 4. **Sixteen oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 14·5 kg with a standard deviation of 0·40 kg. Does the sample mean differ significantly from the intended weight of 16 kg ?**

Solution. We are given : $n = 16$, $\bar{X} = 14{\cdot}5$, $s = 0{\cdot}40$, $\mu = 16$

Let us take the null hypothesis that there is no difference between the sample mean and intended mean *i.e.*,

$$H_0 : \mu = 16 \quad \text{and} \quad H_1 : \mu \neq 16 \quad (\Rightarrow \text{Two tailed test})$$

$$S = \sqrt{\frac{n}{n-1} s^2} = \sqrt{\frac{16}{16-1} \times (\cdot 4)^2} = \sqrt{\frac{2{\cdot}56}{15}} = 0{\cdot}413$$

$$\left[\because s = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n-1}} \right]$$

Applying *t*–test :

$$\therefore \quad |t| = \frac{|\bar{X} - \mu|}{S} \cdot \sqrt{n} = \frac{|14{\cdot}5 - 16|}{0{\cdot}413} \times \sqrt{16} = 14{\cdot}52$$

Degrees of freedom $= v = n - 1 = 16 - 1 = 15$

For $v = 15$, $t_{0{\cdot}05}$ for two tailed test $= 2{\cdot}131$

Since, the calculated value of $|t| = 14{\cdot}52$ > the table value of t, we reject the null hypothesis. It means that the sample mean differ significant from the intended mean 16 kg.

Aliter : The value of can also be calculated using the female;

$$|t| = \frac{|\bar{X} - \mu|}{s} \cdot \sqrt{n-1}$$

$$= \frac{|14{\cdot}5 - 16|}{0{\cdot}4} \times \sqrt{16-1} = \frac{1{\cdot}5 - 3{\cdot}872}{0{\cdot}4} = 14{\cdot}52$$

Example 5. **A consumer testing agency while examining a new automobile for gasoline mileage performance found that 12 readings of miles covered per gallon under the normal conditions resulted in a average of 16 miles per gallon with a standard deviation of 1·8 miles. Do the sample results support the manufacturer's claim that the new automobile gives a performance of more than 15 miles per gallon ? Use $\alpha = 0{\cdot}10$, assuming that the distribution of mileage performance per gallon is approximately normal.**

Solution. We are given : $n=12, \bar{X}=16, s=1\cdot8, \mu=15$

Null hypothesis $H_0 : \mu=15$

Alternative hypothesis $H_1 : \mu>15$ ($\Rightarrow$ Right tailed test)

$$S=\sqrt{\frac{n}{n-1}s^2}=\sqrt{\frac{12}{12-1}\times(1\cdot8)^2}=1\cdot88$$

Applying *t*–test :

$$\therefore \quad t=\frac{\bar{X}-\mu}{S}\cdot\sqrt{n}=\frac{16-15}{1\cdot88}\sqrt{12}=1\cdot84$$

Degrees of freedom $=v=n-1=12-1=11$

For $v=11$, $t_{0\cdot10}$ for one tailed test $=1\cdot363$

Since, the calculated value of $|t|=1\cdot84>$ table vlaue of t at 10% level of significance, the null hypothesis is rejected and conclude that the sample data support the manufacturer's claim of improved mileage performance.

Aliter : The value of t can also be calculated by using the formula :

$$|t|=\frac{|\bar{X}-\mu|}{s}\cdot\sqrt{n-1}$$

$$=\frac{16-15}{1\cdot8}\times\sqrt{12-1}=1\cdot84.$$

EXERCISE – 1

1. Six boys are chosen at random from a school and their heights are found to be in inches : 63, 63, 64, 66, 60, 68. Discuss the suggestion that the mean height in the population is 65 inches ? (Given the table value of t for 5df at 5% level is 2·57).

   [Ans. $t=0\cdot888$, Accept H_0**]**

2. Ten specimens of copper wires drawn from two large lots have the following breaking strength (in kg. wt) :

   578, 572, 570, 568, 572, 578, 570, 572, 596, 584

   Test whether the mean breaking strength of the lot may be taken to be 580 kg. wt.

   [Ans. $t=1\cdot481$, Accept H_0**]**

3. 10 students are selected at random from a college and their marks in Hindi are found to be as follows :

   71, 72, 73, 75, 76, 77, 78, 79, 79, 80

   In the light of the marks, test whether, the average marks in Hindi of the college are 75. (The value of t at 5% for $v=9$ is 2·262) **[Ans.** $t=1\cdot054$, Accept H_0**]**

4. A random sample of 16 values from a normal population showed a mean of 41·5 and the sum of squares of deviations from the mean equal to 135. Can it be assumed that the mean of the population is 43·5. Use 5% level of significance. **[Ans.** $t=2\cdot67$, Reject H_0**]**

5. Ten cartons are taken from an automatic filling machines. The mean weight is 11·802 and standard deviation 0·1502. Does the sample mean differ significanctly from the intended weight 11·6 ? (Given for 9 degree of freedom at 5% level is 2·262).

   [Ans. $t=4\cdot034$, Reject H_0**]**

6. A random sample of size 10 has mean as 40 and standard devaition = 5. Can this sample be regarded as taken from the population having 42 as mean ? (For $v = 9$, $t_{0.05} = 2 \cdot 262$).

[Ans. $t = 1 \cdot 2$, Accept H_0]

7. A soft drink vending machine is set to dispense 8 ounces per cup. If the machine is tested 9 times yielding a mean cup fill of 8·2 ounces with a standard deviation of 0·3 ounces, what can we conclude about the hull hypothesis of $\mu = 8$ ounces against the alternative hypothesis of $\mu > 8$ ounces at $\alpha = 0 \cdot 01$. [Ans. $|t| = 1 \cdot 885$ Accept H_0]

---

(2) Test of hypothesis about diffeence between two means in case of independent samples : Let two independent random samples of sizes n_1 and n_2 ($n_1 < 30$ and $n_2 < 30$) be drawn from two normal populations with means μ_1 and μ_2 and equal standard deviation ($\sigma_1 = \sigma_2 = \sigma_3$). To test whether the two populations means are equal or whether the difference $\bar{X}_1 - \bar{X}_2$ is significant, we use *t*-test and the appropriate test statistic '*t*' to be used is :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Procedure : The following steps are taken while testing the hypothesis about difference between two means :

(1) Set up the null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, there is no significant difference between two populations means :

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

or $H_1 : \mu_1 > \mu_2$ or $\mu_1 < \mu_2$ ($\Rightarrow$ One tailed test)

(2) If the population standard deviation σ_1 and σ_2 are equal *i.e.*, $\sigma_1 = \sigma_2 = \sigma$, the values of *S* is computed by using any of the following formula :

(a) When deviations are taken from actual mean :

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

(b) When actual mean is in fraction and deviations are taken from the assumed mean :

$$S = \sqrt{\frac{\Sigma(X_1 - A_1)^2 + \Sigma(X_2 - A_2)^2 - n_1(\bar{X}_1 - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{n_1 + n_2 - 2}}$$

Where, A_1 and A_2 are the assumed means of two samples

(c) When standard deviations of two samples s_1 and s_2 are given :

$$S = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(3) The values of $\bar{X}_1$, $\bar{X}_2$ n_1 and n_2 and *S* are substituted in the above stated formula.

(4) Degrees of freedom are worked out by using the following formula :

Degrees of freedom $= v = n_1 + n_2 - 2$

The other steps such as (i) level of significance (ii) table value of *t* (iii) decision making for testing the difference between the two means are the same as those given in testing of the hypothesis.

The testing procedure of the differnece of two population means is clarified by the following examples :

Example 6.

In a test given to two groups of students, the marks o

First Group :	18	20	36	50	49
Second Groups :	29	28	26	35	30

Examine the significance of difference between the students of the above two groups. (The value of t at 5

94 Example 7. ...t the null ...s of the

Solution Let us take the hypothesis that there is no significant di... ...the mean marks of the two groups of students *i.e.*,

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

Group I X_1	$\bar{X}_1 = 37$ $(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$	Group II X_2	$\bar{X}_2 = 34$ $(X^2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$
18	−19	361	29	−5	25
20	−17	289	28	−6	36
36	−1	1	26	−8	64
50	+13	169	35	+1	1
49	+12	144	30	−4	16
36	−1	1	44	+10	100
34	−3	9	46	+12	144
49	+12	144			
41	+4	16			
$\Sigma X_1 = 333$ $n_1 = 9$		$\Sigma(X_1 - \bar{X}_1)^2$ $= 1{,}134$	$\Sigma X_2 = 238$ $n_2 = 7$		$\Sigma(X^2 - \bar{X}_2)^2$ $= 386$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{333}{9} = 37, \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{238}{7} = 34$$

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{1134 + 386}{9 + 7 - 2}} = \sqrt{\frac{1520}{14}} = \sqrt{108 \cdot 571} = 10 \cdot 42$$

Applying t–test :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{37 - 34}{10 \cdot 42}\sqrt{\frac{9 \times 7}{9 + 7}} = \frac{3}{10 \cdot 42} \times 1 \cdot 984 = 0 \cdot 571$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 9 + 7 - 2 = 14$

For $v = 14$, $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 14$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that the mean marks of the students of two groups do not differ significantly.

6. A r r **Two independent samples of 8 and 7 items gave the following values :**

Sample A :	9	11	13	11	15	9	12	14
Sample B :	10	12	10	14	9	8	10	

Examine whether the difference between the means of two samples is significant at 5% level.

Solution. Let us take the null hypothesis that there is no significant difference in the two means *i.e.*,

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

X_1	$A_1 = 12$ $(X_1 - A_1)$	$(X_1 - A_1)^2$	X_2	$A_2 = 10$ $(X_2 - A_2)$	$(X_2 - A_2)^2$
9	– 3	9	10	0	0
11	– 1	1	12	+ 2	4
13	+ 1	1	10	0	0
11	– 1	1	14	+ 4	16
15	+ 3	9	9	– 1	1
9	– 3	9	8	– 2	4
12	0	0	10	0	0
14	+ 2	4			
$\Sigma X_1 = 94$ $n_1 = 8$ $\bar{X}_1 = \frac{94}{8}$ $= 11{\cdot}75$		$\Sigma(X_1 - A_1)^2$ $= 34$	$\Sigma X_2 = 73$ $n_2 = 7$ $\bar{X}_2 = \frac{73}{7}$ $= 10{\cdot}43$		$\Sigma(X_2 - A_2)^2$ $= 25$

Since, the actual means are not whole numbers, we take 12 as assumed for X_1 and 10 as assumed for X_2 :

$$S = \sqrt{\frac{\Sigma(X_1 - A_1)^2 + \Sigma(X_2 - A_2)^2 - n_1(\bar{X}_1 - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{34 + 25 - 8(11{\cdot}75 - 12)^2 - 7(10{\cdot}43 - 10)^2}{8 + 7 - 2}}$$

$$= \sqrt{\frac{34 + 25 - 0{\cdot}5 - 1{\cdot}2943}{13}} = \sqrt{4{\cdot}4004} = 2{\cdot}0977$$

Applying t-test,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{11{\cdot}75 - 10{\cdot}43}{2{\cdot}0977} \times \sqrt{\frac{8 \times 7}{8 + 7}}$$

$$= \frac{1.32}{2.0977} \times 1{\cdot}932 = 1{\cdot}2158$$

Degrees of freedom $= v = 8 + 7 - 2 = 13$

For $v = 13$, $t_{0.05}$ for two tailed test $= 2.16$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that there is no significant difference in the means of the two samples.

Aliter : The value of S can also be calculated as:

$$\Sigma (X_1 - \bar{X}_1)^2 = \Sigma d_1^2 - (\Sigma d_1)^2 / n_1 = 34 - (2)^2 / 8 = 33.5$$

$$\Sigma (X_2 - \bar{X}_2)^2 = \Sigma d_2^2 - (\Sigma d_2)^2 / n_2 = 25 - (3)^2 / 7 = 23.71$$

$$S = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2 + \Sigma (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{33.5 + 23.71}{8 + 7 - 2}}$$

$$= \sqrt{\frac{57.21}{13}} = 2.097$$

Example 8. **The mean life of a sample of 10 electric light bulbs was found to be 1456 hours with $s = 423$ hours. A second sample of 17 bulbs chosen from a differnt batch showed a mean life of 1280 hours with $s = 398$ hours. Is there a significant differnce between the means of the two batches ? Here s is the standard deviation.**

Solution. We are given : $n_1 = 10$, $\bar{X}_1 = 1456$, $s_1 = 423$

$n_2 = 17$, $\bar{X}_2 = 1280$, $s_2 = 398$

Let us assume that there is no significant difference between the means of two batches, *i.e.*,

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$S = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(10 - 1) \times 423^2 + (17 - 1) \times 398^2}{10 + 17 - 2}}$$

$$= \sqrt{\frac{1610361 + 2534464}{25}}$$

$$= \sqrt{\frac{4144825}{25}} = 407.17$$

Applying t–test :

$$|t| = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{1456 - 1280}{407.17} \times \sqrt{\frac{10 \times 17}{10 + 17}}$$

$$= \frac{176}{407.17} \times 2.50 = \frac{440}{407.17} = 1.0806$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 10 + 17 - 2 = 25$

For $v = 25$, $t_{.05}$ for two tailed test $= 2 \cdot 06$

Since, the calculated value of $|t| = 1 \cdot 0806 <$ the table value of t, we accept H_0 and conclude that there is no significant difference between the means of two batches.

Example 9. **Two salesmen *A* and *B* are wroking in a certain district. From a sample survey conducted by the Head Office, the following results were obtained. State whether there is any significance difference in the average sales between the two salesmen :**

	A	*B*
No. of sales	20	18
Average	170	205
Standard deviation	20	25

Solution. We are given : $n_1 = 20,\ \bar{X}_1 = 170,\ s_1 = 20 \Rightarrow s_1^2 = 400$

$n_2 = 18,\ \bar{X}_2 = 205,\ s_2 = 25 \Rightarrow s_2^2 = 625$

Let us assume that there is no significant differnce in the average sales between two salesmen *i.e.*, $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

$$S = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(20-1) \times 400 + (18-1) \times 625}{20 + 18 - 2}}$$

$$= \sqrt{\frac{7600 + 10625}{36}} = \sqrt{\frac{18225}{36}} = 22 \cdot 5$$

Applying *t*–test :

$$|t| = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$|t| = \frac{|170 - 205|}{22 \cdot 5} \cdot \sqrt{\frac{20 \times 18}{20 + 18}}$$

$$= \frac{35}{22 \cdot 5} \times 3 \cdot 077 = 4 \cdot 786$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 20 + 18 - 2 = 36$

For $v = 36$, $t_{.05}$ for two tailed test $= 1 \cdot 96$

Since, the calculated value of $|t| = 4 \cdot 786 >$ the table value of t, we reject H_0 and conclude that there is a significant difference in the sales between two salesmen.

Example 10. **On an examination in statistics, 10 students in one class showed a mean grade of 80 with a standard deviation of 8, while 12 students in another class showed a mean grade of 76 with a standard deviation of 10. Using $\alpha = \cdot 01$ level of significance, determine whether the first group is superior to the second group.**

Solution. We are given : $n_1 = 10,\ \bar{X}_1 = 80, s_1 = 8 \Rightarrow s_1^2 = 64$

$n_2 = 12,\ \bar{X}_2 = 76, s_2 = 10 \Rightarrow s_2^2 = 100$

Let us assume that mean grade of two groups don't differ significantly *i.e.*, $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 > \mu_2$ ($\Rightarrow$ Right tailed test)

$$S = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(10-1) \times 64 + (12-1) \times 100}{10 + 12 - 2}}$$

$$= \sqrt{\frac{576 + 1100}{20}} = \sqrt{\frac{1676}{20}} = \sqrt{83 \cdot 8} = 9 \cdot 15$$

Applying *t*–test :

$$|t| = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{80 - 76}{9 \cdot 15} \times \sqrt{\frac{10 \times 12}{10 + 12}}$$

$$= \frac{4}{9 \cdot 15} \times 2 \cdot 33 = 1 \cdot 0185 \approx 1 \cdot 02$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$

For $v = 20$, $t_{.01}$ for one tailed test $= 2 \cdot 528$

Since, the calculated value of $|t|$ < the table value of t, we accept H_0 and conclude that the mean of the two groups are the same *i.e.*, first group is not superior to the second group.

Aliter : This question can also be solved using two–tailed test. Let us have; $H_0 : \mu_1 = \mu_2$ *i.e.*, the first group is not superior to the second group.

And $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

In the light of the facts given, the relevant test statistic 't' computed is :

$$|t| = 1 \cdot 0185 \approx 1 \cdot 02$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$

For $v = 20$, $t_{.01}$ for two tailed test $= 2 \cdot 845$.

Since, the calculated value of $|t|$ < table value of t, we accept H_0 and conclude that the first group is not superior to the second group.

Example 11. **The means of two random samples of sizes 9 and 7 are 196·42 and 198·82 respectively. The sum of the squares of the deviations from the means are 26·94 and 18·73 respectively. Can the samples be considered to have been drawn from the same normal population ?**

Solution. Given : $n_1 = 9,\ \bar{X}_1 = 196 \cdot 42,\quad \Sigma (X_1 - \bar{X}_1)^2 = 26 \cdot 94$

$n_2 = 7,\ \bar{X}_2 = 198 \cdot 82,\quad \Sigma (X_2 - \bar{X}_2)^2 = 18 \cdot 73$

Let us take the null hypothesis that the samples are drawn from the same normal population *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_1 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{26 \cdot 94 + 18 \cdot 73}{9 + 7 - 2}} = \sqrt{\frac{45 \cdot 67}{14}} = \sqrt{3 \cdot 26} = 1 \cdot 81$$

Applying *t*–test :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$|t| = \frac{|196 \cdot 42 - 198 \cdot 82|}{1 \cdot 81} \sqrt{\frac{9 \times 7}{9 + 7}}$$

$$= \frac{2 \cdot 40 \times 1 \cdot 98}{1 \cdot 81} = 2 \cdot 6254$$

Degrees of freedom $(v) = n_1 + n_2 - 2 = 9 + 7 - 2 = 14$

For $v = 14$, $t_{.05}$ for two tailed test $= 2 \cdot 145$

Since, the calculated value of t is greater than the table vlaue, we reject the null hypothesis and conclude that the difference is significant. Thus, both the samples have not been taken from same population.

Example 12. **A random sample of seven week old chickens reared on a high protein diet weight : 12, 15, 11, 16, 14, 14 and 16 ounces, another random sample of five chickens similarly treated except that they received a low protein diet weight : 8, 10, 14, 10 and 13 ounces. Test whether there is significant evidence that additional protein has increased the weight of chickens.**

(The table value of t for 10 degrees of freedom at 5% level of significance is 2.228)

Solution. Let us take the hypothesis that additional protein has not increased the weight of chickens, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

X_1	$\bar{X}_1 = 14$ $(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$	X_2	$\bar{X}_2 = 11$ $(X_2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$
12	– 2	4	8	– 3	9
15	+ 1	1	10	– 1	1
11	– 3	9	14	+ 3	9
16	+ 2	4	10	– 1	1
14	0	0	13	+ 2	4
14	0	0			
16	+ 2	4			
$\Sigma X_1 = 98$ $n_1 = 7$		$\Sigma(X_1 - \bar{X}_1)^2 = 22$	$\Sigma X_2 = 55$ $n_2 = 5$		$\Sigma(X_2 - \bar{X}_2)^2 = 24$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{98}{7} = 14 \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{55}{5} = 11$$

$$S = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2 + \Sigma (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{22 + 24}{7 + 5 - 2}} = \sqrt{\frac{46}{10}} = 2 \cdot 14$$

Applying *t*–test,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{14 - 11}{2 \cdot 14} \times \sqrt{\frac{7 \times 5}{7 + 5}}$$

$$= \frac{3}{2 \cdot 14} \times 1 \cdot 708 = 2 \cdot 394$$

Degrees of freedom $(v) = n_1 + n_2 - 2 = 7 + 5 - 2 = 10$

For $v = 10$, $t_{.05}$ for two tailed test $= 2 \cdot 228$

Since, the calculated value of t is more than the table value, the null hypothesis is rejected. Hence, there is a significant evidence that additional protein has increased the weight of chickens.

Aliter : This question can also be solved by using one tailed test.

Let us have, $H_0 : \mu_1 = \mu_2$ *i.e.*, additional protein has not increased the weight of chickens.

And $H_1 : \mu_1 > \mu_2$ (as we want to conclude that additional protein has increased the weight)

It is a one tailed test.

In the light of the facts given, the relevant test statistic computed is :

$$t = 2 \cdot 394$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 7 + 5 - 2 = 10$

For $v = 10$, $t_{.05}$ for one tailed test $= 1 \cdot 812$.

Since, the calculated vlaue of $|t| = 2 \cdot 394$ is more than the table value, the null hypothesis is rejected. Hence, there is a significant evidence that additional protein has increased the weight of chickens.

EXERCISE – 2

1. The height of six randomly chosen soldiers are in inches : 76, 70, 68, 69, 69 and 68. Those of 6 randomly chosen sailors are 68, 64, 65, 69, 72 and 64. Discuss in the light of these data the suggestion that soldiers are on the average taller than sailors. Use *t*–test. (Table value of *t* at 5% level for 10 d.f. = 2·23) [Ans. $t = 1 \cdot 66$, Accept H_0]
2. Below are given the gain in weights (lbs) of lions fed on two diets x_1 and x_2 :

   Gain in weights (ibs)

Diet x_1 :	25	32	30	32	24	14	32			
Diet x_2 :	24	34	22	30	42	31	40	30	32	35

Test at 5% level whether the two diets differ as regards their effect of the mean increase in weight. [**Ans.** $t = 1{\cdot}585$, Reject H_0]

3. Two kinds of fertilizers were applied to 15 plots. Other conditions remaining the same, the yields in quintals are given below :

Fertilizer I :	18	31	28	22	26	40	45	
Fertilizer II :	20	14	48	40	44	34	32	30

Examine the significance of difference between the mean yields due to different kinds of fertilizers. [**Ans.** $t = 0{\cdot}535$, Accept H_0]

4. Test the sgnificance of the difference of means of the two samples at 5% level of significance from the following data :

	No. of Items	Mean	S.D.
Sample A :	6	40	8·0
Sample B :	5	50	10·0

(The table value of t for 9 d.f. at 5% level is 2·262) [**Ans.** $t = 1{\cdot}84$, Accept H_0]

5. Test the significance of the difference of the means of two samples at 5% level of significance from the following data :

	Sample Size	Mean	Variance
Sample A :	10	1000	100
Sample B :	12	1020	121

(The table value of t for 20 d.f. at 5% level is 2·086) [**Ans.** $t = 4{\cdot}42$, Reject H_0]

6. Two salesmen A and B are working in a certain district. From a sample survery conducted by the head office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

	A	B
No. of sales	10	18
Average sales (Rs.)	170	205
Standard deviation (in Rs.)	20	25

[**Ans.** $t = 3{\cdot}79$, Reject H_0]

7. In a sample of 10 electric lamps the mean life and S.D. was found to be 1456 hours and 423 hours respectively. Another 17 rnadomly selected lamps had a mean life of 1200 hours and S.D. 398 hours. Is there any significant difference in their mean values ?

[**Ans.** $t = 1{\cdot}578$, Accept H_0]

8. Two salesmen are working in a shop. The number of items sold by them in a week are given as :

S_1	25	32	30	32	24	14	34
S_2	24	34	22	30	42	31	40

Test at 5% level whether the performance of two salesmen differ significantly.

9. Measurement performed on random samples of two kinds of cigarettes yielded the following results on their nicotine content (in mgs.) : ($t_{0.05} = 2.179$ at df = 12)

Brand A :	21.4	23.6	24.8	22.4	26.3
Brand B :	22.4	27.7	23.5	29.1	25.8

Assuming that nicotine content is distributed normally, test the hypothesis that Brand B has a higher nicotine content than Brand A.

(**Hint** : Use one tailed Test)

[**Ans.** $|t| = 1 \cdot 315$, Accept H_0]

10. The nicotine contents in milligrams of two samples of tobacco were found to be as follows :

Sample A :	24	27	26	21	25	
Sample B :	27	30	28	31	22	26

Can it be said that two samples come from normal population having the same mean ?

[**Ans.** $|t| = 1 \cdot 92$, Accept H_0]

---

(3) Test of hypothesis about diffeence between two means with dependent samples (*i.e.*, paired data) or paired *t*-test : *t*–test is also used to test the hypothesis about difference between two means in case of paired data *i.e.*, when the sample items are the same but different situations are beings analysed. For example, performance of some students are taken down before and after extra coaching and we want to find the effectiveness of extra coaching. To test the significance of the difference between two means in two situations in case of paired data, the appropriate test statistic '*t*' to be used is :

$$t = \frac{\bar{d}}{S} \cdot \sqrt{n}$$

Where $\bar{d}$ = mean of the difference, *i.e.*, $\frac{\Sigma d}{n}$, n = size of the sample

S = standard deviation of the difference.

Procedure : The following steps are taken while testing the hypothesis of difference between two means in case of paired data :

(i) Set up the null hypothesis $H_0 : \bar{d} = 0$ or $\mu_1 = \mu_2$

Alternative hypothesis $H_1 : \bar{d} \neq 0$ or $\mu_1 \neq \mu_2$ (two tailed test)

or $H_1 : \bar{d} > 0$ *i.e.*, $\mu_1 < \mu_2$ or $\mu_2 > \mu_1$ (one tailed test)

(ii) Find the difference between each matched pair as :

$d = I - II$ or $II - I$

(iii) Calculate the mean of the difference as :

$$\bar{d} = \frac{\Sigma d}{n}$$

(iv) The value of the standard deviation of the difference is computed by using the following formula :

$$S = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$$

or

$$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$$

(4) The values of $\bar{d}$, S and n are substituted in the above stated formula.

(5) Degrees of freedom are worked by using the following formula :

Degrees of freedom $= v = n - 1$

The other steps such as (i) level of significance, (ii) table value of t (iii) decision making for testing the difference between the two means are the same as given in test of hypothesis.

Example 13. **Ten students of M. Com. were given a test in the Business Statistics. They were imparted a month's special coaching and a second test was conducted at the end of it. The results were as follows :**

Students	Marks in Ist Test	Marks in IInd Test
1	36	32
2	40	42
3	38	30
4	36	24
5	42	18
6	38	64
7	40	32
8	46	40
9	58	52
10	62	38

Do the marks give an evidence that the students have benefited by extra coaching ?

(Given : $v = d.f. = 9$, $t_{.05} = 2 \cdot 202$)

Solution. This is a problem on paired observations from two dependent samples. Here the paired t-test is used.

Computation of Mean and S.D.

Students	Marks in Ist Test X_1	Marks in IInd Test (after extra coaching) X_2	$d = X_2 - X_1$	d^2
1	36	32	– 4	16
2	40	42	2	4
3	38	30	– 8	64
4	36	24	– 12	144
5	42	18	– 24	576

6	38	64	26	676
7	40	32	−8	64
8	46	40	−6	36
9	58	52	−6	36
10	62	38	−24	576
$n=10$			$\Sigma d=-64$	$\Sigma d^2=2192$

$$\text{Mean difference} = \bar{d} = \frac{\Sigma d}{n} = \frac{-64}{10} = -6{\cdot}4$$

$$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}} = \sqrt{\frac{2192 - (-6{\cdot}4)^2 \times 10}{10-1}}$$

$$= \sqrt{\frac{2192 - 409{\cdot}6}{9}} = 14{\cdot}07$$

Let us have the null hypothesis $H_0 : \bar{d} = 0$ *i.e.*, $\mu_2 - \mu_1 = 0$ (*i.e.*, students are not benefited by extra coaching).

Alternative Hypothesis $H_1 : \bar{d} > 0$ *i.e.*, $\mu_2 - \mu_1 > 0 \Rightarrow \mu_2 > \mu_1$ (*i.e.*, students are benefited by extra coaching)

Here one tailed test is to be applied

Aliter : The value of S can also be calculated as :

$$\Sigma (d - \bar{d})^2 = \Sigma d^2 - (\Sigma d)^2 / n = 2192 - (-64)^2 / 10$$

$$= 1782{\cdot}4$$

$$S = \sqrt{\frac{\Sigma (d - \bar{d})^2}{n-1}} = \sqrt{\frac{178{\cdot}4}{9}} = 14{\cdot}07$$

Applying *t*–test :

$$|t| = \frac{\bar{d}}{S} \cdot \sqrt{n}$$

$$= \frac{-6{\cdot}4}{14{\cdot}07} \times \sqrt{10} = \frac{20{\cdot}238}{14{\cdot}07} = 1{\cdot}438$$

Degrees of freedom $= v = n - 1 = 10 - 1 = 9$

For $v = 9$, $t_{.05}$ for one tailed test $= 1{\cdot}833$

Since the calculated value of $|t| = 1{\cdot}438$ in less than the table value of t, we accept H_0 and conclude that extra coaching has not benefited the students.

Aliter : This question can also be solved using two tailed test :

Let us take the hypothesis that there is no difference in the marks before and after special coaching *i.e.*, extra coaching has not benefitted the students *i.e.*,

$$H_0 : \bar{d} = 0$$

And $H_1 : \bar{d} \neq 0$ ($\Rightarrow$ Two tailed test)

In the light of the facts given, the relevant test statistics 't' computed is

$$|t| = 1 \cdot 438$$

For $v = 9$, $t_{.05}$ for two tailed test $= 2 \cdot 262$

Sincet he calculated value of $|t| = 1 \cdot 438$ is less than the table value of t, we accept H_0 and conclude that extra coaching has not benefited the students.

Example 14. **A certain medicine given to each of the 9 patients resulted in the following increase in blood pressure :**

7, 3, −1, 4, −3, 5, 6, −4, −1

Can it be concluded that the medicine will, in general, be accompained by an increase in blood pressure ? (Given $t_{.05}(8) = 2 \cdot 0306$)

Solution. Let us assume that the medicine does not increase the blood pressure *i.e.*,

$H_0 : \bar{d} = 0$ *i.e.*, $\mu_1 = \mu_2$ and $H_1 = \bar{d} \neq 0$ ($\Rightarrow$ Two tailed test)

d	7	3	−1	4	−3	5	6	−4	−1	$\Sigma d = 16$
d^2	49	9	1	16	9	25	36	16	1	$\Sigma d^2 = 162$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{16}{9} = 1 \cdot 778$$

$$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}} = \sqrt{\frac{162 - (1 \cdot 778)^2 \times 9}{9-1}} = 4 \cdot 086$$

$$t = \frac{\bar{d}}{S} \cdot \sqrt{n} = \frac{1 \cdot 778}{4 \cdot 086} \times \sqrt{9} = 1 \cdot 30$$

Degrees of freedom $(v) = n - 1 = 9 - 1 = 8$.

For $v = 8$ d.f., $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 0306$

Since, the calculated value of $|t| = 1 \cdot 30$ is less than the table value of t, we accept the null hypothesis and conclude that the medicine in general does not increase the blood pressure.

Aliter : This question can also be solved using one tailed test.

$H_0 : \bar{d} \leq 0$ *i.e.*, there is no increase in blood pressure after the medicine.

$H_1 : \bar{d} > 0$ ($\Rightarrow$ One tailed test)

In the light of the facts given, the relevant test-statistics 't' computed is :

$$|t| = \frac{\bar{d}}{S} \cdot \sqrt{n} = 1 \cdot 30$$

For d.f. $= v = 8$, $t_{.05}$ for one tailed test $= 1 \cdot 86$

Since, the calculated value of $|t| = 1 \cdot 30$ is less than the table value of t, we accept H_0 and conclude that medicine in general does not increase the blood pressure.

Aliter : The value S can also be calculated on :

$$\Sigma (d - \bar{d})^2 = \Sigma d^2 - (\Sigma d)^2 \ / \ n = 162 - (16)^2 \ / \ 9 = 133 \cdot 55$$

$$S=\sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}=\sqrt{\frac{133\cdot55}{9-1}}=4\cdot085$$

Example 15. 10 prsons were appointed in a clerical position in an office. Their performance were noted giving a test and the mark recorded out of 50. They were given 6 months training and again they were given a test and marks were recorded out of 50.

Employees :	A	B	C	D	E	F	G	H	I	J
Before taining :	25	20	35	15	42	28	26	44	35	48
After training :	26	20	34	13	43	40	29	41	36	46

By applying the *t*–test, can it be concluded that the employees have been benefitted by the training ? (Given for d.f. = 9, $t_{.05}=2\cdot262$)

Solution. Let us take the hypothesis that the employees have not been benefitted by the training *i.e.*, $H_0:\bar{d}=0$ and $H_1:\bar{d}\neq0$. ($\Rightarrow$ two tailed test)

Before Training (I)	After Training (II)	d (II – I)	d^2
25	26	+1	1
20	20	0	0
35	34	–1	1
15	13	–2	4
42	43	+1	1
28	40	+12	144
26	29	+3	9
44	41	–3	9
35	36	+1	1
48	46	–2	4
$n=10$		$\Sigma d=10$	$\Sigma d^2=174$

$$\bar{d}=\frac{\Sigma d}{n}=\frac{10}{10}=1$$

$$S=\sqrt{\frac{\Sigma d^2-(\bar{d})^2\times n}{n-1}}$$

$$=\sqrt{\frac{174-10(1)^2}{10-1}}=4\cdot269$$

Applying *t*–test :

$$t=\frac{\bar{d}}{S}\cdot\sqrt{n}=\frac{1}{4\cdot269}\times\sqrt{10}=\frac{3\cdot1692}{4\cdot269}=0\cdot741$$

Degrees of freedom $=v=n-1=10-1=9$

For $v = 9$, $t_{.05}$ for two tailed test $= 2 \cdot 262$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that the employees have not been benefitted from the training.

Aliter : The value S can also be calculated on :

$$\Sigma (d - \bar{d})^2 = \Sigma d^2 - (\Sigma d)^2 / n = 174 - (10)^2 / 10$$

$$S = \sqrt{\frac{\Sigma (d - \bar{d})^2}{n-1}} = \sqrt{\frac{164}{10-1}} = 4 \cdot 269$$

EXERCISE – 3

1. A certain stimulus administered to each of 12 patients resulted in the following increase in blood pressure : 5, 2, 8, –1, 3, 0, – 2, 1, 5, 0, 4 and 6. Can it be concluded that the stimulus, will in general be accompanied by an increase in blood pressure ? (For $v = 11$, $t_{0 \cdot 01} = 2 \cdot 21$)

[**Ans.** $t = 2 \cdot 90$, Reject H_0]

2. 12 Students were given intensive coaching and two tests were conducted, first test before coaching and second test after coaching. The scores of two tests are given below :

Student	1	2	3	4	5	6	7	8	9	10	11	12
Marks in 1st test	50	42	51	26	35	42	60	41	70	55	62	38
Marks in 2nd test	62	40	61	35	30	52	68	51	84	63	72	50

Has the coaching helped in improving the scores ?

[**Ans.** $t = 4 \cdot 885$ Reject H_0, Coaching is careful]

3. The sales data of an item in six shops before and after a special promotional compaign are as under :

Shops :	A	B	C	D	E	F
Before campaign :	53	28	31	48	50	42
After campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success ? Test at 5% level of significance.

[**Ans.** $t = 2 \cdot 78$, Reject H_0]

4. An I.Q. (Intelligence Quotient) test was conducted for 5 officers before and after a training. The reuslts are given below :

Officer :	I	II	III	IV	V
I.Q. before training :	110	120	123	132	125
I.Q. after training :	120	118	125	136	121

Test whether there is any change in I.Q. after the training programme.
(Give $t_{.01}(4) = 4 \cdot 604$)

[**Ans.** $t = \cdot 817$, Accept H_0]

5. Eight students were given Test in statistics, and after one month's coaching, they were given another test of the similar nature. The following table gives the difference in their marks in the second Test over the first :

Roll Number :	1	2	3	4	5	6	7	8
Difference in Marks :	4	– 2	6	– 8	12	5	– 7	2

Is the difference in marks statistically significant ?

[**Ans.** $t = 0 \cdot 623$, Accept H_0 and training is not effective *i.e.* difference in marks is statistically significant]

---

(4) Test of Hypothesis about Coefficient of Correlation : Suppose that a random sample (x_i, y_i) of size n has been drawn from a bivariate normal population and let r be the observed sample correlation coefficient. To test the hypothesis that the correlation coefficient in the population is zero *i.e.*, $\rho = 0$ or the observed correlation is significant or not, we use t–test and the appropriate test statistic t to be used is

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$$

Where, r = sample correlation coefficient, n = size of the sample.

Procedure : The following steps are taken while testing the hypothesis about coefficient of correlation :

(1) Set up the null hypothesis $H_0 : \rho = 0$, *i.e.*, correlation coefficient in the population is zero or the observed correlation coefficient is not significant.
Alternative hypothesis $H_1 : \rho \neq 0$ ($\Rightarrow$ Two tailed test)

(2) Substituting the values of r and n in the above stated formula.

(3) Degrees of freedom are worked out by using the following formula :
Degrees of freedom $= v = n - 2$

The other steps such as (i) level of significance, (ii) table value of t and (iii) decision making for testing the population correlation coefficient are the same as those given for the test of hypothesis.

Example 16. **A correlation coefficient of 0·6 is discovered in a sample of 18 observations. Is it significant at 1% level ?**

Solution. Given, $r = 0 \cdot 6, n = 18$

$H_0 : \rho = 0$

and $H_1 : \rho \neq 0$ (Two tailed test)

Applying t–test :

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2} = \frac{0 \cdot 6}{\sqrt{1-(\cdot 6)^2}} \times \sqrt{18-2} = \frac{\cdot 6}{\cdot 8} \times 4 = 3$$

Degrees of freedom $= v = n - 2 = 18 - 2 = 16$

For $v = 16$, $t_{.05}$ for two tailed test $= 2 \cdot 92$

Since the calculated value of $|t| = 3$ is greater than the table value of t, we reject H_0 and conclude that the observed value of the correlation coefficient is significant.

Example 17. **A random sample of 27 observations from a normal populations gives a correlation coefficient of – 0·4. Is this significant of the existence of correlation in the population ?**

(Given for $v = 25$, $t_{.01} = 2 \cdot 79$)

Solution. We are given : $n = 27, r = -0 \cdot 4$

$H_0 : \rho = 0$

And $H_1 : \rho \neq 0$ ($\Rightarrow$ Two tailed test)

Aplying t–test,

$$t=\frac{r}{\sqrt{1-r^2}}\cdot\sqrt{n-2}$$

$$=\frac{0\cdot4}{\sqrt{1-(-\cdot4)^2}}\sqrt{27-2}$$

$$=\frac{\cdot4}{\cdot917}\times5=2\cdot18$$

Degres of freedom $=v=n-1=27-2=25$

For $v=25$, $t_{\cdot05}$ for two tailed test $=2\cdot79$

Since the calculated value of $|t|$ is less than the table value, we accept H_0 and conclude that the correlation of the sample is not significant to warrant the existence of such correlation in the population.

Example 18. **(a) How many pairs of observations must be included in a sample in order that an observed correlation coefficient of value 0·42 shall have a calculated value of t greater than 2·72?**

(b) Find the least value of r in a random sample of 27 pairs of values from a bivariate population which would be significant at 5% level.

Solution. (a) Given : $r=0\cdot42$, Critical value, 2·72

We have $$t=\frac{r}{\sqrt{1-r^2}}\times\sqrt{n-2}$$

In order that the calculated value of t may be greater than 2·72,

$$\frac{r}{\sqrt{1-r^2}}\times\sqrt{n-2}>2\cdot72 \quad\Rightarrow\quad \frac{0\cdot4}{\sqrt{1-(\cdot42)^2}}\times\sqrt{n-2}>2\cdot72$$

$$\Rightarrow\quad \frac{0\cdot42\sqrt{n-2}}{\sqrt{1-\cdot1764}}>2\cdot72 \quad\Rightarrow\quad \frac{0\cdot42\sqrt{n-2}}{\sqrt{\cdot8236}}>2\cdot72$$

$$\Rightarrow\quad \frac{0\cdot42\sqrt{n-2}}{0\cdot908}>2\cdot72 \quad\Rightarrow\quad 0\cdot4625\sqrt{n-2}>2\cdot72$$

$$\Rightarrow\quad \sqrt{n-2}>\frac{2\cdot72}{0\cdot4625} \quad\Rightarrow\quad \sqrt{n-2}>5\cdot88\Rightarrow n-2>34\cdot57$$

or $n>34\cdot57+2$ or $36\cdot57$ (Squaring both sides)

Hence, the required value of n would be greater than 36·57 *i.e.*, 37 at least.

(b) In order that the 'r' may be significant, the calculated value of t must be more than the table value of r for 25 d.f. at 5% for two tailed test.

We have $$t=\frac{r}{\sqrt{1-r^2}}\times\sqrt{n-2}$$ and $t_{\cdot05}(25)=2\cdot06$

Thus, according to given condition, $t>2\cdot06$

$$\frac{5r}{\sqrt{1-r^2}}>2\cdot06 \quad\text{or}\quad 5r>2\cdot06\sqrt{1-r^2}$$

or $25r^2 > 4 \cdot 2436(1-r^2)$ or $25r^2 > 4 \cdot 2436 - 4 \cdot 2436\, r^2$

$\Rightarrow$ $29 \cdot 2436\, r^2 > 4 \cdot 2436$ or $r^2 > \dfrac{4 \cdot 2436}{29 \cdot 2436}$

$\Rightarrow$ $r > 0 \cdot 38$

Hence, any value of r which is more than $0 \cdot 38$ would be significant for 25 d.f. at 5% level.

EXERCISE – 4

1. A random sample of 18 pairs from a normal population showed a correlation coefficient of $0 \cdot 4$. Is this value significant of correlation in the population ?
   [Ans. $t = 1 \cdot 76$, Accept H_0]
2. Find the least value of r in a sample of 18 pairs a bivariate normal population significant at 5% level. [Ans. $|r| = \cdot 468$]
3. How many pairs of observations must be included in a sample in order that an observed correlation coefficient of value $0 \cdot 52$ shall have a calculated value of t greater than $2 \cdot 82$?
   [Ans. $n = 24$]
4. A random sample of 25 from a normal universe gives correlation coefficient of $-0 \cdot 48$. Is this significant of the existance of correlation in the population ?
   [Ans. $|t| = 2 \cdot 624$, Reject H_0]
5. A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of $0 \cdot 58$. Is this significant of correlation in the population ?
   [Ans. $t = 3 \cdot 56$, Reject H_0]
6. A study of the heights of 25 pairs of husbands and their wives in a factory shows that the coefficient of correlation is $0 \cdot 37$. Test whether correlation is significant or not (The value of t at 5% for 23 degrees of freedom is $2 \cdot 069$) [Ans. $t = 1 \cdot 903$, Accept H_0]
7. Is a value of $r = -0 \cdot 48$ significant if obtained from a sample of 25 pairs of values from a normal population ? [Ans. $t = 2 \cdot 624$, Reject H_0]

(2) Fisher's Z—Transformation

t–test is used to test the significance of the correlation coefficient if the value of the population correlation coefficient is zero. If the population correlation coefficient is other than zero or difference between two sample correlation coeffiicents are to be tested, then the t-rest can not used and in that case Fisher Z-test is used. Fisher's Z-test has two applications :

(1) To test whether an observed value of r differs significantly from some hyptothetical value of population correlation coefficient, other than zero, the appropriate test statistic to used is :

$$|Z| = \frac{Z_r - Z_\rho}{SE_z}$$

Procedure

Its testing procedure is as follows ;

(i) Set up the null hypothesis $H_0 : \rho = \rho_0$ *i.e.*, population correlation coefficient is equal to a specified value or there is no difference between r and ρ.

(ii) Thereafter, the value of r (sample correlation coefficient) and ρ (population correlation coefficient) are changed into Z–transformation by using the following formula :

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1\cdot1513\log_{10}\left(\frac{1+r}{1-r}\right)$$

(iii) **Z–transformation of r**

$$Z_s = 1\cdot1513 \cdot \log_{10}\left(\frac{1+r}{1-r}\right)$$

Z–transformation of ρ

$$Z_\rho = 1\cdot1513 \cdot \log_{10}\left(\frac{1+\rho}{1-\rho}\right)$$

Note : If ρ is not known, then it is taken as zero in which case $\rho = 0$.

(iv) The standard error of Z is worked out as under :

$$S.E._z = \frac{1}{\sqrt{n-3}}$$

(v) Finally, we compute the value of Z as follows :

$$Z = \frac{Z_s - Z_p}{\frac{1}{\sqrt{n-3}}} = (Z_s - Z_p) \times \sqrt{n-3}$$

(vi) The calculated value of Z is compared with 1·96 at 5% level of significance and 2·58 at 1% level of significance.

If $|Z| > 1\cdot96$, the difference is considered significant at 5% level of significance otherwise insignificiant.

If $|Z| > 2\cdot58$, the difference is considered significant at 1% level of significance otherwise insignificant.

The following examples illustrate the application of Z-test.

Example 19. **Test the significance of the coefficient of correlation $r = 0\cdot5$ discovered in a sample of 19 paired of observation against hypothetical correlation $\rho = 0\cdot7$. Apply Fisher's Z—test.**

Solution. Let us take the hypothesis that correlation coefficient in the population is 0·7 *i.e.*,

$H_0 : \rho = 0\cdot7$ and $H_1 : \rho \neq 0\cdot7$

Given : $r = 0\cdot5$, $\rho = 0\cdot7$

Applying Z–transformation, we obtain :

Z–transformation of r

$$Z_s = 1\cdot1513\log_{10}\left(\frac{1+r}{1-r}\right)$$

$$= 1\cdot1513 \cdot \log_{10}\left(\frac{1+0\cdot5}{1-0\cdot5}\right)$$

$$= 1\cdot1513\log_{10}\left(\frac{1\cdot5}{0\cdot5}\right)$$

$$= 1\cdot1513 \times \log 3$$

$$= 1\cdot1513 \times \cdot4771 = 0\cdot549$$

Z–transformation of ρ

$$Z_\rho = 1\cdot1513\log_{10}\left(\frac{1+\rho}{1-\rho}\right)$$

$$= 1\cdot1513 \cdot \log_{10}\left(\frac{1+0\cdot7}{1-0\cdot7}\right)$$

$$= 1\cdot1513\log_{10}\left(\frac{1\cdot7}{0\cdot3}\right)$$

$$= 1\cdot1513 \times \log(5\cdot67)$$

$$= 1\cdot1513 \times 0\cdot7536 = 0\cdot868$$

Applying Fisher's Z–test

$$\therefore \quad |Z| = \frac{Z_s - Z_p}{\frac{1}{\sqrt{n-3}}} = \frac{|0.549 - 0.868|}{\frac{1}{\sqrt{19-3}}}$$

$$= \frac{0.319}{1} \times \sqrt{16} = 0.319 \times 4 = 1.276$$

Since, the calculated value of $|Z|$ is less than 1·96, we accept the null hypothesis at 5% level and conclude that the population correlation coefficient is 0·7.

(2) Testing the significance of the difference between two independent sample correlation coeffiicents.

Z–test can also be used to test the significance of the difference of the two sample correlation coefficients.

Procedure

Its testing procedure is as follows :

(i) First of all, set up the null hypothesis that there is no difference between two correlation coefficients *i.e.*, $H_0 : \rho_1 = \rho_2$

(iii) Thereafter, the two values r_1 and r_2 are changed into Z–transformation by using the following formula :

$$Z = 1.1513 \log_{10}\left(\frac{1+r}{1-r}\right)$$

(iv) The standard error of the difference between Z_1 and Z_2 is worked out as under :

$$S.E_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

(v) Finally, we compute the value of $|Z|$ as follows :

$$|Z| = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Where $Z_1 = 1.1513 \log_{10}\left(\frac{1+r_1}{1-r_1}\right)$

$$Z_2 = 1.1513 \log_{10}\left(\frac{1+r_2}{1-r_2}\right)$$

(v) If the calculated value of $|Z|$ is greater than 1·96 at 5% level of significane, the difference between two r's is significant.

Example 20. **The following data give sample sizes and correlation coefficient. Test the significance of the difference between the values using Fisher's Z–test.**

Sample size	Value of r
5	0·87
12	0·56

Solution : Let us take the hypothesis that two correlation coefficients do not differ significantly *i.e.*, $H_0 : \rho_1 = \rho_2$ and $H_1 : \rho_1 \neq \rho_2$ ($\Rightarrow$ Two tailed test)

Z–transformation of r_1

$$Z_1 = 1{\cdot}1513 \log_{10}\left(\frac{1+r_1}{1-r_1}\right)$$

$$= 1{\cdot}1513 \log_{10}\left(\frac{1+0{\cdot}87}{1-0{\cdot}87}\right)$$

$$= 1{\cdot}1513 \log_{10}\left(\frac{1{\cdot}87}{0{\cdot}13}\right)$$

$$= 1{\cdot}1513 \log_{10}(14{\cdot}384)$$

$$= 1{\cdot}1513 \times 1{\cdot}1577 = 1{\cdot}33$$

Z–transformation of r_2

$$Z_2 = 1{\cdot}1513 \log_{10}\left(\frac{1+r_2}{1-r_2}\right)$$

$$= 1{\cdot}1513 \log_{10}\left(\frac{1+0{\cdot}56}{1-0{\cdot}56}\right)$$

$$= 1{\cdot}1513 \log_{10}\left(\frac{1{\cdot}56}{0{\cdot}44}\right)$$

$$= 1{\cdot}1513 \times \log_{10} 3{\cdot}545$$

$$= 1{\cdot}1513 \times 0{\cdot}5495 = 0{\cdot}63$$

Applying Fisher's Z–test,

$$|Z| = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{1{\cdot}33 - 0{\cdot}63}{\sqrt{\frac{1}{5-3} + \frac{1}{12-3}}} = \frac{0{\cdot}70}{\sqrt{0{\cdot}61}} = \frac{0{\cdot}70}{0{\cdot}78} = 0{\cdot}9$$

As the value of Z is less than 1·96 at 5% level of significance, we accept the null hypothesis and conclude that the difference is not significant.

EXERCISE – 5

1. Test the significance of the correlation $r = +0{\cdot}75$ from a sample of size 30 against hypothetical correlation $\rho = 0{\cdot}55$. [Ans. $|Z| = 1{\cdot}8$, Accept H_0]
2. From a sample of 10 pair of observations the correlation is 0·5 and the corresponding population value is 0·3. Is this difference significant at 5% level of significance ? Apply Z–test. [Ans. $|Z| = 0{\cdot}96$, Accept H_0]
3. The following data give sample sizes and correlation coefficients Test the significance of the difference between two values using Fisher's Z–test.

Sample size	Value of r
23	0·87
28	0·56

[Ans. $|Z| = 1.$, Accept H_0]

4. A correlation coeffcient of 0·6 is obtained from a sample of 19 paired observations. Is it significantly different from 0·4 ? [Ans. $|Z| = 1{\cdot}076$, Accept H_0]

(3) F–Test (Variance Ratio Test)

F–test is named after the greater statistician **R.A. Fisher.** F–test is used to test whether the two independent estimates of population variance differ significantly or whether the two samples may be regarded as drawn from the normal population having the same variance. For carrying out the test, we calculate F–statistic. F–statistic is defined as :

$$F = \frac{\text{Larger estimate of population variance}}{\text{Smaller estimate of population variance}} = \frac{S_1^2}{S_2^2} \quad \text{where,} \quad S_1^2 > S_2^2$$

Procedure

Its testing procedure is as follows :

(i) Set up null hypothesis that the two population variances are equal *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2$

(ii) The variances of the random samples are calculated by using formula :

$$S_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} \quad \text{or} \quad \frac{n_1}{n_1 - 1} s_1^2 \quad \text{or} \quad \frac{1}{n_1 - 1}\left[\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right]$$

$$S_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_1 - 1} \quad \text{or} \quad \frac{n_2}{n_2 - 1} s_2^2 \quad \text{or} \quad \frac{1}{n_2 - 1}\left[\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right]$$

(iii) The variance ratio F is computed as :

$$F = \frac{S_1^2}{S_2^2} \qquad \text{where, } S_1^2 > S_2^2$$

(iv) The degrees of freedom are computed. The degrees of freedom of the larger estimate of the population variance is denoted by v_1 and the smaller estimate by v_2. That is,

v_1 = degrees of freedom for sample having larger variance $= n_1 - 1$

v_2 = degrees of freedom for sample having smaller variance $= n_2 - 1$

(v) Then from the F–table given at the end of the book, the value of F is found for v_1 and v_2 with 5% level of significance.

(vi) Then we compare the calculated value of F with the table value of $F_{.05}$ for v_1 and v_2 degrees of freedom. If the calculated value of F exceeds the table value of F, we reject the null hypothesis and conclude that the differnce between the two variances is significant. On the other hand, if the calculated value of F is less than the table value, the null hypothesis is accepted and conclude that both the samples have come from the population having same varaince.

The following examples illustrate the applications of F–test :

Example 21. **In a sample of 8 observations, the sum of squared deviations of items from the mean was 94·5. In another sample of 10 observations, the value was found to be 101·7 Test whether the difference is significant at 5% level. (You are given that at 5% level of significance, the critical value of F for $v_1 = 7$ and $v_2 = 9$ df is 3·29).**

Solution. Let us take the hypothesis that the difference in the variances of the two samples is not significant *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2$

We are given : $n_1 = 8,\ \Sigma(X_1 - \bar{X}_1)^2 = 94{\cdot}5$

$$n_2 = 10,\ \Sigma(X_2 - \bar{X}_2)^2 = 101{\cdot}7$$

$$S_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{94{\cdot}5}{8 - 1} = \frac{94{\cdot}5}{7} = 13{\cdot}5$$

$$S_2^2 = \frac{\Sigma(X_2 - \bar{X}_1)^2}{n_1 - 1} = \frac{101{\cdot}7}{10 - 1} = \frac{101{\cdot}7}{9} = 11{\cdot}3$$

Applying F–test,

$$F = \frac{S_1^2}{S_2^2} = \frac{13 \cdot 5}{11 \cdot 3} = 1 \cdot 195$$

For $v_1 = 8 - 1 = 7$ and $v_2 = 10 - 1 = 9$, $F_{.05} = 3 \cdot 29$

The calculated value of F is less than the table value. Hence, we accept the null hypothesis and conclude that the difference in the varainces of two samples is not significant at 5% level.

Example 22. **Two random samples drawn from normal populations are :**

Sample I :	20	16	26	27	23	22	18	24	25	19		
Sample II :	27	33	42	35	32	34	38	28	41	43.	30	37

Obtain estimates of the variances of the two populations and test whether two populations have the same variances.

Solution. Let us take the hypothesis that two populations have the same variance *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_0^2$

Sample I X_1	$(X_1 - \bar{X}_1)$ $\bar{X}_1 = 22$	$(X_1 - \bar{X}_1)^2$	Sample II X_2	$(X_2 - \bar{X}_2)$ $\bar{X}_2 = 35$	$(X_2 - \bar{X}_2)^2$
20	– 2	4	27	– 8	64
16	– 6	36	33	– 2	4
26	+ 4	16	42	+ 7	49
27	+ 5	25	35	0	0
23	+ 1	1	32	– 3	9
22	0	0	34	– 1	1
18	– 4	16	38	+ 3	9
24	+ 2	4	28	– 7	49
25	+ 3	9	41	+ 6	36
19	– 3	9	43	+ 8	64
			30	– 5	25
			37	+ 2	4
$\Sigma X_1 = 220$ $n_1 = 10$		$\Sigma (X_1 - \bar{X}_1)^2 = 120$	$\Sigma X_2 = 420$ $n_2 = 12$		$\Sigma (X_2 - \bar{X}_2)^2 = 314$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{220}{10} = 22; \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{420}{12} = 35$$

$$S_1^2 = \frac{\Sigma (X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10 - 1} = \frac{120}{9} = 13 \cdot 33$$

$$S_2^2 = \frac{\Sigma (X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12 - 1} = \frac{314}{11} = 28 \cdot 545$$

Applying F–test, we have

$$F = \frac{S_2^2}{S_1^2} = \frac{28 \cdot 545}{13 \cdot 33} = 2 \cdot 14 \qquad \text{where } S_2^2 > S_1^2$$

For $v_1 = 11$ and $v_2 = 9$, $F_{.05} = 3 \cdot 11$

Since, the calculated value of F is less than the table value, the null hypothesis is accepted and hence it may be concluded that the two populations have the same variance.

Example 23. **The following data relate to a random sample of Government employees in two states of Indian Union :**

	State I	State II
Sample size :	16	25
Mean monthly income (Rs.)	440	460
Sample variance	40	42

In the light of the data, test the hypothesis that the varainces of two populations are equal.

Solution. Let us take the null hypothesis that the variances of the two populations are equal *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2$

We are given : $n_1 = 16 \quad s_1^2 = 40$

$$n_2 = 25 \quad s_2^2 = 42$$

$$S_1^2 = \frac{n_1}{n_1 - 1} \cdot s_1^2 = \frac{16}{16-1} \times 40 = \frac{16}{15} \times 40 = \frac{640}{15} = 42 \cdot 67$$

$$S_2^2 = \frac{n_2}{n_2 - 1} \cdot s_2^2 = \frac{25}{25-1} \times 42 = \frac{25}{24} \times 42 = \frac{1050}{24} = 43 \cdot 75$$

$$F = \frac{43 \cdot 75}{42 \cdot 67} = 1 \cdot 025 \qquad \text{where, } S_2^2 > S_1^2$$

For $v_1 = 24$, and $v_2 = 15$, $F_{.05} = 2 \cdot 29$

Since, the calculated value of F is less than the table value of F, we accept the null hypothesis and hence it may be concluded that the variances of two populations are equal.

Example 24. **Two independent samples of 8 and 7 items respectively had the following values of variable (weight in grams) :**

Sample I :	9	11	13	11	15	9	12	14
Sample II :	10	12	10	14	9	8	10	

Do the two estimates of population variance differ significantly ?

Solution. Let us take the null hypothesis that the two populations have the same variance *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2$.

Sample I		Sample II	
X_1	X_1^2	X_2	X_2^2
9	81	10	100
11	121	12	144
13	169	10	100
11	121	14	196
15	225	9	81
9	81	8	64
12	144	10	100
14	196		
$\Sigma X_1 = 94$, $n_1 = 8$	$\Sigma X_1^2 = 1138$	$\Sigma X_2 = 73$, $n_2 = 7$	$\Sigma X_2^2 = 785$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{94}{8} = 11{\cdot}75; \qquad \bar{X}_2 = \frac{X_2}{n_2} = \frac{73}{7} = 10{\cdot}43$$

Since, the actual means are in fractions, we make use of original values. Thus,

$$S_1^2 = \frac{1}{n_1 - 1}\left[\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right] = \frac{1}{8-1}\left[1138 - \frac{(94)^2}{8}\right] = \frac{33{\cdot}5}{7} = 4{\cdot}78$$

$$S_2^2 = \frac{1}{n_2 - 1}\left[\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right] = \frac{1}{7-1}\left[785 - \frac{(73)^2}{7}\right] = \frac{23{\cdot}7}{6} = 3{\cdot}95$$

Applying F–test, we have :

$$F = \frac{4{\cdot}78}{3{\cdot}95} = 1{\cdot}21$$

For $v_1 = 7$ and $v_2 = 6$, $F_{.05} = 4{\cdot}21$

Since, the calculated value of F is less than the table value of F, we accept the null hypothesis and it may be concluded that the two estimates of population variances do not differ significantly.

AN IMPORTANT TYPICAL EXAMPLE

Example 25. **Can the following two samples be regarded as coming from the same normal population ?**

Sample	Size	Sample mean	Sum of squares of deviation from mean
1	10	12	120
2	12	15	314

Solution. To test if two independent samples have been drawn from the same normal population, we have to test (i) the equality of population means, and (ii) the equality of population variances.

Equality of means will be tested by applying t–test and equality of variance will be tested by F–test. Since, t–test assumes $\sigma_1^2 = \sigma_2^2$, we first apply F–test and then t–test.

***F*–test :** Set up $H_0 : \sigma_1^2 = \sigma_2^2$

We are given : $n_1 = 10,\ \Sigma(X_1 - \bar{X}_2)^2 = 314$

$$S_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10-1} = \frac{120}{9} = 13 \cdot 33$$

$$S_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12-1} = \frac{314}{11} = 28 \cdot 55$$

Applying F–test,

$$F = \frac{S_2^2}{S_1^2} = \frac{28 \cdot 55}{13 \cdot 33} = 2 \cdot 14$$

For $v_1 = 11 - 1 = 10$ and $v_2 = 10 - 1 = 9$, $F_{.05} = 3 \cdot 14$

The calculated values of F is less than the table value. Hence, we accept the null hypothesis and conclude that the difference in the varainces of two samples is not significant at 5% level.

Since, $\sigma_1^2 = \sigma_2^2$, we can now apply t–test for testing $H_0 : \mu_1 = \mu_2$

t–test : Set up $H_0 : \mu_1 = \mu_2$

We are given : $n_1 = 10$ $\bar{X}_1 = 12$ $\Sigma(X_1 - \bar{X}_1)^2 = 120$

$n_2 = 12$ $\bar{X}_2 = 15$ $\Sigma(X_2 - X_2)^2 = 314$

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{120 + 314}{10 + 12 - 2}} = \sqrt{\frac{434}{20}} = \sqrt{21 \cdot 7} = 4 \cdot 65$$

Aplying t–test, we have

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{12 - 15}{4 \cdot 65} \times \sqrt{\frac{10 \times 12}{10 + 12}} = 1 \cdot 506$$

Degrees of freedom $(v) = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$

Table value of t for 20 d.f. at 5% level of signficance $= 2 \cdot 086$.

Since, the calculated value of t is less the table value, we accept the null hypothesis and conclude that the difference in means is not significant.

Hence, we may regard that the given samples to have been drawn from same population.

EXERCISE – 6

1. Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 inches squares. Calculate the value of F and say whether it is significant or not at 5% level of significance ? (Given $F_{.05}$ for 8 and 7 d.f = 3·73)

OR

Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 inches squares. Can they be regarded as drawn from the same normal population at $\alpha = 0.05$? [Ans. $F = 1.54$, the samples can be regarded as drawn from the same normal population]

2. Two random samples drawn from normal populations are :

Sample I :	66	67	75	76	82	84	88	90	92		
Sample II :	64	66	74	78	82	85	87	92	93	95	97

Obtain estimates of the variances of the two populations and test whether the two populations have the same variances. (Given $F = 3.35$ at 5% level for $v_1 = 10$ and $v_2 = 8$)

[Ans. $F = 1.414$, the two populations have the same variance)

3. For a random sample of 10 pigs fed on diet *A*, the increase in weight in pounds in certain periods were :

10, 6, 16, 17, 13, 12, 8, 14, 15, 9

For another random sample of 12 pigs fed on diet *B*, the increase in weight in the same period were :

7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17

Test whether both the samples come from population having same variance.
(Given : $F_{.05}$ for $v_2 = 11$, $v_2 = 9$ is 3·112).

[Ans. $F = 2.14$, samples come from population having same variance]

4. It is known that the mean diameters of rivets produced by two firms *A* and *B* are practically the same but standard deviations differ. For 22 rivets produced by firm *A*, the standard deviation is 2·9 mm, while for 16 rivets manufactured by firm *B*, the standard deviation is 3·8 mm. Compute the statistic you would use to test whether the product of firm *A* has the same variability as those of firm *B*.

[Ans. $F = 1.748$, the two populations have the same variance]

5. In a test given to two groups of students drawn from two normal populations, the marks obtained were as follows :

Group A :	18	20	36	50	49	36	34	39	41
Group B :	29	28	26	35	30	44	46		

Examine at 5% level, whether the two populations have the same variance.

[Ans. $F = 2.103$, populations have the same variances]

6. Two sets of random samples drawn from normal population are given below. Obtain the estimates of the variances of the two populations and test whether the two populations have the same variance. Use F–test.

Sample I :	20	16	26	27	23	22	18	24	25	19	30	37
Sample II :	27	33	42	35	32	34	38	28	41	43		

(Table value of F for $v_1 = 11$ and $v_2 = 9$ at 5% level = 3·112)

[Ans. $F = 2.142$, population have the same variance]

7. The variability in the tensile strength of two types of steel wire is to compared. Given a sample of 10 observations of type *A* wire yielding a variance of 100·4 and a sample of 12 observations of type *B* wire yielding a variance of 115·5, test the hypothesis that the two populations have equal variances.

[Ans. $F = 1.0625$, population have the same variance]

8. Two samples were drawn independently from two normal populations. The summary statistics are :

$n_1 = 8,\ \Sigma(X_1 - \bar{X}_1)^2 = 84{\cdot}4$ inches

$n_2 = 13,\ \Sigma(X_2 - \bar{X}_2)^2 = 102{\cdot}6$ inches

In the light of the data, test whether the two variances differ significantly.

[Ans. $F = 1{\cdot}410$, variances do not differ]

9. Can the following two samples be regarded as coming from the same normal population ?

Sample	Size	Sample mean	Sum of squares of deviation from mean
1	10	15	90
2	12	14	108

[Hints : Use F–test; $F = \frac{S_1^2}{S_2^2} = 1.019$ and t–test for $H_0 : \mu_1 = \mu_2, |t| = 0{\cdot}742$]

[Ans. The samples are drawn from the same normal population]

10. The means of two random samples of size 9 and 7 are 196·42 and 198·82 respectively. The sum of the squares of the deviations from the mean are 26·94 and 18·73 respectively. Can the samples by considered to have been drawn from the same normal population ?

[Ans. $F = 1{\cdot}078, |t| = 2{\cdot}634$, Reject H_0]

11. The following data relate to a random sample of Government employees in two states of Indian Union. First carry out a test of hypothesis that the variance of the two populations are equal. In the light of the result of the above test, carry out a test of hypothesis that the means of two populations are equal :

	State I	State II
Sample Size	16	25
Mean monthly income of sample employees (in days.)	440	460
Sample Variance	40	42

[Ans. $F = 1{\cdot}025$, Accept H_0 , $t = 9{\cdot}72$, Accept H_0]

MISCELLANEOUS SOLVED EXAMPLES

Example 26. **Prices of shares of a company on different days in a month were found to be :**

66, 65, 69, 70, 69, 71, 70, 63, 64 and 68

Discuss whether the mean price of the shares should be 65.

(The table value of t for 9 degree of freedom at 5% level is 2·262)

Solution. Let us take the hypothesis that the mean price of the share is 65, *i.e.*,

$H_0 : \mu = 65$ and $H_1 : \mu \neq 65$ (⇒ Two tailed test)

X	$A = 67$ $d = X - A$	d^2
66	−1	1
65	−2	4
69	+2	4

70	+ 3	9
69	+ 2	4
71	+ 4	16
70	+ 3	9
63	– 4	16
64	– 3	9
68	+ 1	1
$n = 10, \Sigma X = 675$	$\Sigma d = 5$	$\Sigma d^2 = 73$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{675}{10} \quad 67 \cdot 5$$

Since, the actual means of X is in fraction, we should take deviations from assumed mean to simplify the calculations.

$$\bar{d} = \frac{\Sigma d}{n} = \frac{5}{10} = 0 \cdot 5$$

$$S = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{73 - 10(0 \cdot 5)^2}{9}} = 2 \cdot 799$$

Applying t–test,

$$t = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n} = \frac{67 \cdot 5 - 65}{2 \cdot 799} \times \sqrt{10} = \frac{2 \cdot 5 \times 3 \cdot 162}{2 \cdot 799} = 2 \cdot 82$$

Degrees of freedom $(v) = n - 1 = 10 - 1 = 9$

For $v = 9$, $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 262$

Since, the calculated value of t is greater than the table vlaue, we reject the null hypothesis and therefore, conclude that mean price of the shares could not be equal to Rs. 65.

Example 27. **To compare the price of a certain commodity in two towns, ten shops were selected at random in each town. The following figures give the price found.**

Town A :	61	62	56	63	56	63	59	56	44	60
Town B :	55	54	47	59	51	61	57	54	64	58

Test whether the average price can be said to be the same in two towns.

Solution. Let us take the hypothesis that there is no difference in the average price of two towns : *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

X_1	$X_1 - \bar{X}_2$	$(X_1 - \bar{X}_1)^2$	X_2	$X^2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
61	3	9	55	– 1	1
62	4	16	54	– 2	4
56	– 2	4	47	– 9	81
63	5	25	59	3	9
56	– 2	4	51	– 5	25
63	5	25	61	5	25

59 56 44 60	1 −2 −14 2	1 4 196 4	57 54 64 58	1 −2 8 2	1 4 64 2
$\Sigma X_1 = 580$ $n_1 = 10$	0	$\Sigma (X_1 - \bar{X}_1)^2$ $= 288$	$\Sigma X_2 = 560$ $n_2 = 10$	0	$\Sigma (X_2 - \bar{X}_2)^2$ $= 216$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{580}{10} = 58, \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{560}{10} = 56$$

$$S = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2 + \Sigma (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{288 + 216}{10 + 10 - 2}} = \sqrt{\frac{504}{18}} = \sqrt{28} = 5 \cdot 29$$

Applying t–test,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{58 - 56}{5 \cdot 29} \sqrt{\frac{10 \times 10}{10 + 10}} = \frac{2 \times 2 \cdot 236}{5 \cdot 29} = 0 \cdot 845$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$

For $v = 18$, $t_{0 \cdot 05} = 2 \cdot 101$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that there is no significant difference in the mean price.

Example 28. **An I.Q. test was administrated to 5 officers before and after they were trained. The results are given below :**

Candidates :	I	II	III	IV	V
I.Q. before training :	110	120	123	132	125
I.Q. after training :	120	118	125	136	121

Test whether there is any change in I.Q. after the training programme. [For $v = 4$, $t_{0 \cdot 01} = 4 \cdot 6$]

Solution. Let us take the hypothesis is that there is no change in I.Q. after the training programme. *i.e.*, $H_0 : \bar{d} = 0$ or $\mu_2 - \mu_1 = 0$ and $H_1 : \bar{d} > 0$ or $\mu_2 - \mu_1 > 0$

($\Rightarrow$ One tailed test)

I.Q. Before	I.Q. After II	d (II – I)	d^2
110	120	+ 10	100
120	118	− 2	4
123	125	+ 2	4
132	136	+ 4	16
125	121	− 4	16
$n = 5$		$\Sigma d = 10$	$\Sigma d^2 = 140$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{10}{5} = 2$$

$$S = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}}$$

$$= \sqrt{\frac{140 - 5(2)^2}{5-1}} = \sqrt{\frac{140-20}{4}} = \sqrt{\frac{120}{4}} = \sqrt{30} = 5 \cdot 477$$

Applying t–test,

$$t = \frac{\bar{d}}{S}\sqrt{n} = \frac{2\sqrt{5}}{5 \cdot 477} = 0 \cdot 817$$

For $v = 4$, $t_{0 \cdot 01} = 4 \cdot 6$

The calculated value of t is less than the table value. We accept the null hypothesis and hence there is no change in I.Q. after the raining programme.

Example 29. **Two types of batteries are tested for their length of life and the following data are obtained :**

	No. of samples	Mean life in hours	Variance
Type A :	9	600	121
Type B :	8	640	144

Is there a significant difference in two mean ? Value of t for 15 degrees of freedom at 5% level is 2·131.

Solution. Let us take the hypothesis that there is no significant difference in mean life of two types of batteries *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

Given : $n_1 = 9$, $\bar{X}_1 = 600$, $s_1^2 = 121$

$n_2 = 8$, $\bar{X}_2 = 640$, $s_2^2 = 144$

$$S = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(9-1) \times 121 + (8-1) \times 144}{9 + 8 - 2}} = \sqrt{\frac{968 + 1008}{15}}$$

$$= \sqrt{\frac{1976}{15}} = \sqrt{131 \cdot 73} = 11 \cdot 47$$

Applying t–test,

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{|600 - 640|}{11 \cdot 47} \times \sqrt{\frac{9 \times 8}{9 + 8}}$$

$$= \frac{40}{11 \cdot 47} \times 2 \cdot 057 = \frac{82 \cdot 28}{11 \cdot 47} = 7 \cdot 17$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 9 + 8 - 2 = 15$

For $v = 15$, $t_{0.05} = 2.131$

Since, the calculated value of t is greater than the table value, we reject H_0 and hence, the difference in the means is significant.

Example 30. **Two types of drugs were used on 5 and 7 patients for reducing their weights. Drug A was imported and drug B indigenous. The decreases in the weight after using drugs for six months are as follows :**

Drug A :	10	12	13	11	14		
Drug B :	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs ? If not, which drug should you buy.

(For $v = 10$, $t_{0.05} = 2.223$)

Solution : Let us take the hypothesis that there is no significant difference in the efficacy of two drugs, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test).

X_1	$\bar{X}_1 = 12$	$(X_1 - \bar{X}_1)^2$	X_2	$\bar{X}_2 - 11$	$(X_2 - \bar{X}_2)^2$
10	– 2	4	8	– 3	9
12	0	0	9	– 2	4
13	+ 1	1	12	+ 1	1
11	– 1	1	14	+ 3	9
14	+ 2	4	15	+ 4	16
			10	– 1	1
			9	+ 2	4
$\Sigma X_1 = 60$ $n_1 = 5$		$\Sigma (X_1 - \bar{X}_1)^2 = 10$	$\Sigma X_2 = 77$ $n_2 = 7$		$\Sigma (X_2 - \bar{X}_2)^2 = 44$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{60}{5} = 12, \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{77}{7} = 11$$

$$S = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2 + \Sigma (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{10 + 44}{5 + 7 - 2}} = \sqrt{\frac{54}{10}} = \sqrt{5 \cdot 4} = 2 \cdot 324$$

Applying t–test,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{12 - 11}{2 \cdot 324} \cdot \sqrt{\frac{5 \times 7}{5 + 7}}$$

$$= \frac{1 \times 1 \cdot 708}{2 \cdot 324} = \frac{1 \cdot 708}{2 \cdot 324} = 0 \cdot 735$$

Degrees of freedom $(v) = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$

For $v = 10$, $t_{0.05} = 2.228$

Since, the calculated value of t is less than the table vlaue, we accept H_0 and conclude that there is no significant difference in the efficacy of two drugs. Since,

drug *B* is indigenous and there is no difference in the efficacy of imported and indigenous drug, we should buy indigenous drug *B*.

Example 31. **Below are given the gain in weights (lbs) of cows fed on two diets *X* and *Y*:**

Gain Weight (lbs)

Diet *X*:	25	32	30	32	24	14	32			
Diet *Y*:	24	34	22	30	42	31	40	30	32	35

Test at 5% level, whether the two diets differ as regards their effect on mean increase in weight (Table value of *t* for 15 degrees of freedom at 5% = 2·131).

Solution. Let us take the null hypothesis that diet *X* and *Y* do not differ significantly with regard to their effect on increase in weight, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$
(⇒ Two tailed test)

X	$\bar{X} = 27$ $X - \bar{X}$	$(X - \bar{X})^2$	Y	$\bar{Y} = 32$ $Y - \bar{Y}$	$(Y - \bar{Y})^2$
25	− 2	4	24	− 8	64
32	+ 5	25	34	+ 2	4
30	+ 3	9	22	− 10	100
32	+ 5	25	30	− 2	4
24	− 3	9	42	+ 10	100
14	− 13	169	31	− 1	1
32	+ 5	25	40	+ 8	64
			30	− 2	4
			32	0	0
			35	+ 3	9
$\Sigma X = 189$ $n_1 = 7$	$\Sigma (X - \bar{X}) = 0$	$\Sigma (X - \bar{X})^2$ $= 266$	$\Sigma Y = 320$ $n_2 = 10$	$\Sigma (Y - \bar{Y}) = 0$	$\Sigma (Y - \bar{Y})^2$ $= 350$

$$\bar{X} = \frac{\Sigma X}{n_1} = \frac{189}{7} = 27, \qquad \bar{Y} = \frac{\Sigma Y}{n_2} = \frac{320}{10} = 32$$

$$S = \sqrt{\frac{\Sigma (X - \bar{X})^2 + \Sigma (Y - \bar{Y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{266 + 350}{7 + 10 - 2}} = \sqrt{\frac{616}{15}} = \sqrt{41 \cdot 066} = 6 \cdot 40$$

Applying *t*–test,

$$|t| = \frac{\bar{X} - \bar{Y}}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{|27 - 32|}{6 \cdot 40} \times \sqrt{\frac{7 \times 10}{7 + 10}}$$

$$= \frac{5}{6 \cdot 40} \times 2 \cdot 029 = \frac{10 \cdot 1459}{6 \cdot 40} = 1 \cdot 58$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 7 + 10 - 2 = 15$

For $v = 15$, $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 131$

Since, the calculated value of *t* is less than the table value, we accept the null hypothesis and conclude that diets *X* and *Y* do not differ significantly as regards their effects on increase in weight is concerned.

Example 32. **A random sample of 18 pairs from a bivariate normal population showed a correlation coefficient of 0·4. Is this value significant of a correlation in the population ?**

Solution. Let us take the hypothesis that the variables are uncorrelated in the population.

$H_0 : \rho = 0$ and $H_1 = \rho \neq 0$ (⇒ two tailed test)

Applying t–test,

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{\cdot}4 \times \sqrt{18-2}}{\sqrt{1-0{\cdot}16}}$$

$$= \frac{0{\cdot}4 \times 4}{\sqrt{0{\cdot}84}} = \frac{1{\cdot}6}{0{\cdot}91} = 1{\cdot}76$$

Degrees of freedom $= v = n-2 = 18-2 = 16$

For $v = 16$, $t_{0{\cdot}05}$ for two tailed test $= 2{\cdot}12$

Since, the claculated value of t is less than the table value, we accept H_0 and hence, the given value of r is not significant.

Example 33. **A correlation coefficient of 0·63 is obtained from a sample of 20 paired of observations. Is it significantly different from 0·5?**

Solution. We are given : $n = 20, r = 0{\cdot}72, \rho = 0{\cdot}8$

Let us take the null hypothesis that the correlation in the population is 0·5 *i.e.*,

$H_0 : \rho = 0{\cdot}5$ and $H_1 : \rho \neq 0{\cdot}5$ (⇒ two tailed test)

Z—transformation or r

$$Z_1 = 1{\cdot}1513 \log_{10} \frac{1+r}{1-r}$$

$$= 1{\cdot}1513 \log_{10} \frac{1+{\cdot}63}{1-{\cdot}63}$$

$$= 1{\cdot}1513 \log 4{\cdot}4$$

$$= 1{\cdot}1513 \times {\cdot}6435 = 0{\cdot}741$$

Z—transformation of ρ

$$Z_2 = 1{\cdot}1513 \log_{10} \frac{1+r}{1-r}$$

$$= 1{\cdot}1513 \log_{10} \frac{1+{\cdot}5}{1-{\cdot}5}$$

$$= 1{\cdot}1513 \log 3$$

$$= 1{\cdot}1513 \times {\cdot}4771 = 0{\cdot}549$$

$$SE_{Z_1 - Z_2} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{20-3}} = \frac{1}{\sqrt{17}} = 0{\cdot}243$$

Applying Fisher's Z—test

$$|Z| = \frac{Z_r - Z_p}{SE_{Z_1 - Z_2}} = \frac{0{\cdot}741 - 0{\cdot}549}{0{\cdot}243} = \frac{0{\cdot}192}{0{\cdot}243} = 0{\cdot}79.$$

The critical value of Z at 5% for two tailed test $= 1{\cdot}96$.

Since, the calculated value of $|Z|$ is less than 1·96 (5%), we accept null hypothesis and conclude that the given correlation coefficient is not significantly different from 0·5.

Example 34. **In a laboratory experiment, two samples gave the following results :**

Sample	Size	Sample Mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test the equality of sample variances at 5% level of significance.

Solution. Let the null hypothesis be that the two population variances are equal *i.e.*,

$$H_0 : \sigma_1^2 = \sigma_2^2$$

We are given :

$n_1 = 10$, $\Sigma(X_1 - \bar{X}_1)^2 = 90$

$n_2 = 12$, $\Sigma(X_2 - \bar{X}_2)^2 = 108$

$$S_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{90}{10-1} = \frac{90}{9} = 10$$

$$S_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{108}{12-1} = \frac{108}{11} = 9 \cdot 82$$

Applying *F*–test,

$$F = \frac{S_1^2}{S_2^2} \qquad \text{where } S_1^2 > S_2^2$$

$$= \frac{10}{9 \cdot 82} = 1 \cdot 018$$

For $v_1 = 10 - 1 = 9$ and $v_2 = 12 - 1 = 11$, $F_{05} = 2 \cdot 90$

Since, the calculated value of F is less than the table value, we accept the null hypothesis and conclude that the two populations have the same variance.

Example 35. **The profit of an automobile dealer varies from day to day. However, the dealer believes that the per day profit averages at least Rs. 3500. The profits per day during the past week were reported to be Rs. 2000, Rs. 3000, Rs. 5200, Rs. 3400, Rs. 2500 and 3700. Would you agree with the belief of the dealer. Use at 0·05 level of significance ?**

Solution. Let us take the hypothesis that the average profit of the dealer is at least Rs. 3500 *i.e.*, $H_0 : \mu \geq 3500$ and $H_1 : \mu < 3500$

[Since, the dealer belief would be false if the average rate is less than 3500]

It is one tailed test.

Sales X	$\bar{X} = 3300$ $(X - \bar{X})$	$d = (X - \bar{X})/100$	d^2
2000	– 1300	– 13	169
3000	– 300	– 3	9
5200	+ 1900	+ 19	361
3400	+ 1000	+ 1	1

2500 3700	– 800 + 400	– 8 + 4	64 16
$\Sigma X = 19800$ $n = 6$		$\Sigma d = 0$	$\Sigma d^2 = 620$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{19800}{6} = 3300$$

$$S = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n-1}} \times c \qquad [\text{here, } (c = 100)]$$

$$= \sqrt{\frac{620}{5}} \times 100 = \sqrt{124} \times 100 = 11 \cdot 1355 \times 100 = 1113 \cdot 55$$

Applying t–test,

$$|t| = \frac{\bar{X} - \mu}{S} . \sqrt{n}$$

$$= \frac{|3300 - 3500|}{1113 \cdot 35} \sqrt{6} = \frac{200}{1113 \cdot 35} \times 2 \cdot 449 = \frac{489 \cdot 897}{1113 \cdot 35} = 0 \cdot 44$$

Degrees of freedom $= v = n - 1 = 6 - 1 = 5$

For $v = 5$, $t_{0 \cdot 05}$ for one tailed test $= 2 \cdot 015$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that the claim of the dealer is justified.

IMPORTANT FORMULAE

t-test

1. Tests of Hypothesis about population mean :

$$t = \frac{\bar{X} - \mu}{S} . \sqrt{n}$$

where, $$S = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n-1}} \text{ or } \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$$

d.f. $= n - 1$

2. Test of Hypothesis about the difference between two population means in case of independent samples :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} . \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Where,

$$S = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2 + \Sigma (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$$

d.f. $= n_1 + n_2 - 2$

3. **Test of Hypothesis about the difference of the two population means with depende samples**

$$t = \frac{\bar{d}}{S} \cdot \sqrt{n}$$

where, $$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$$

$$\text{d.f.} = v = n - 1$$

4. **Test of Hypothesis about correlation coefficient :**

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

where, $\text{d.f.} = v = n - 2$

Fisher's Z-test

5. **Test of Hypothesis about correlation coefficient ($H_0 : \rho = \rho_0$) :**

$$|Z| = \frac{Z_r - Z_p}{SE_z}$$

6. **Test of Hypothesis about two correlation coefficients ($H_0 : \rho_1 = \rho_2$) :**

$$|Z| = \frac{Z_S - Z_p}{SE_{Z_1 - Z_2}}$$

F-test

8. **Test of Hypothesis about two population variances ($H_0 : \sigma_1^2 = \rho_2^2$) :**

$$F = \frac{S_1^2}{S_2^2} \quad \text{where, } S_1^2 > S_2^2$$

QUESTIONS

1. Define student's *t*-test and explain some of its applications.
2. Explain how *t*–test is used to test the significance of the difference between the mean two samples.
3. Explain briefly various application of the *t*–test.
4. Explain how *t*–test is used to test the significance of the sample correlation coefficient sample drawn from a bivariate normal population.
5. Discuss Fisher's Z-test for testing the significance of correlation coefficient.
6. Discuss the *F*–test for testing the equality of two sample variances.
7. Discuss the usefulness of *F*–test.
8. Explain the procedure for testing hypothesis regarding equaltiy of two variances.
9. Explain how would test the significance of the correlation coefficient in case of s sample.

❀❀❀

5

Chi-Square Test

INTRODUCTION

The Chi-Square test (χ^2-test) is an important test amongst several tests of significance developed by the statisticians is. Chi-Square, symbolically written as χ^2 (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for testing the significance of population variance. As a non-parametric test, it can be used as a test of goodness of fit and as a test of attributes. Thus, the Chi-Square test is applicable to a very large number of problems in practice which can be summed up under the following heads :

(1) χ^2-test as a test for population variance.

(2) χ^2-test as a non-parametric test.

Let us discuss them briefly

(1) χ^2-test as a test for population variance : χ^2-test is often used to test the significance of population variance i.e. we can use this test to judge if a random sample has been draw from a normal population with mean (μ) and with a specified variance (σ_0^2).

PROCEDURE :

(i) Set up the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 > \sigma_0^2$

(ii) We compute χ^2 by using any one of the following formula :

$$\chi^2 = \frac{\Sigma(x - \bar{x})^2}{\sigma^2} \quad \text{or} \quad \frac{ns^2}{\sigma^2}$$

Where $$s^2 = \frac{\Sigma(x - \bar{x})^2}{n}$$

$\Rightarrow$ $$ns^2 = \Sigma(x - \bar{x})^2$$

(iii) No. of degrees of freedom are worked out by using the following formula :
Degrees of freedom = v = n - 1

(iv) Obtain the table value of χ^2 with reference to the degrees of freedom for the given problem and the desired level of significance.

(v) If the calculated value of χ^2 > tabulated value of χ^2, we reject the null hypothesis H_0.
Otherwise, we accept H_0.

Note : In case H_1 is two tailed test, the procedure is slightly different. Accept H_0 if the calculated $\chi^2 < \chi^2_{\alpha/2}$ or calculated $\chi^2 < \chi^2_{1-\alpha/2}$ i.e. $\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{1-\alpha/2}$ and reject H_0 if calculated $\chi^2 > \chi^2_{\alpha/2}$ or calculated $\chi^2 < \chi^2_{1-\alpha/2}$.

Example 1 : **With variance of a normal population $\sigma^2 = 5.80$ and the sum of the squares of the deviations of 15 sample values from their mean is 150, compute the χ^2-value.**

Solution. We are given : $n = 15, \sigma^2 = 5.80$ and $\Sigma(X - \bar{X})^2 = 150$

The χ^2-value is calculated as :

$$\chi^2 = \frac{\Sigma(X - \bar{X})^2}{\sigma^2} = \frac{150}{5.8} = 25.86$$

Example 2 : **A sample of 10 units drawn from a normally distributed population shows a variance $s^2 = 25$. Test the hypothesis that the population variance $\sigma^2 = 36$ using $\alpha = 0.01$ level of significance.**

Solution. We are given : $n = 10, s^2 = 25, \sigma^2 = 36, \alpha = .01$

Let us take the null hypothesis that the population variance is 36 i.e.

$H_0 : \sigma^2 = 36$ and $H_1 : \sigma^2 > 36$ ($\Rightarrow$ Right tailed test)

The value of χ^2 is calculated as follows:

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$= \frac{10 \times 25}{36} = 6.94$$

Degrees of freedom $= v = n - 1 = 10 - 1 = 9$

For $v = 9, \chi^2_{.01} = 21.7$

Since, the calculated vlaue of χ^2 is less than the table value of $\chi^2_{0.01}$ we accept H_0 and conclude that the population variance $\sigma^2 = 36$.

Example 3 : **A random sample of size 20 from of normal population gives a sample mean of 42 and sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9. Clearly state the alternative hypothesis you allow for and the level of significance adopted.**

Solution. We are given : $n = 20, \bar{X} = 42, s = 6 \Rightarrow s^2 = 36, \sigma = 9 \Rightarrow \sigma^2 = 81$

Let us take the null hypothesis that the population standard deviation is 6, i.e.

$H_0 : \sigma = 9$ and $H_1 : \sigma > 9$ ($\Rightarrow$ Right Tailed Test)

χ^2-value is calculated as :

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$= \frac{20 \times 36}{81} = 8.89$$

Degrees of freedom $= v = 20 - 1 = 19$

For $v = 19$, $\chi^2_{.05} = 30.1$

Since, the calculated vlaue of χ^2 is less than the table value of $\chi^2_{.05}$, we accept H_0 and conclude that the population standard deviation is 9.

ple 4 : Weights in kg of 10 students are given below :

38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that variance of the distribution of weight of all students from which the above sample of 10 students was drawn, is equal to 20 kgs ? Test this at 5% and 1% level of significance. (At 9 d.f., $\chi^2_{.05} = 16.92$, $\chi^2_{.01} = 21.67$ at 10 d.f; $\chi^2_{.05} = 18.31$, $\chi^2_{.01} = 23.21$)

ion. Let us take the null hypothesis that population variance is 20, *i.e.*

$H_0 : \sigma^2 = 20$ and $H_1 : \sigma^2 > 20$ ($\Rightarrow$ Right tailed test)

Applying χ^2- test

X	$\bar{X} = 47$ $X - \bar{X}$	$(X - \bar{X})^2$
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	64
48	1	1
52	5	25
49	2	4
$\Sigma X = 470$ $n = 10$		$\Sigma(X - \bar{X})^2 = 280$

$$\therefore \quad \bar{X} = \frac{470}{10} = 47$$

χ^2-value is calculated as :

$$\chi^2 = \frac{\Sigma(X - \bar{X})^2}{\sigma^2} = \frac{280}{20} = 14$$

Degrees freedom $= v = n - 1 = 10 - 1 = 9$

For $v=9, \chi^2_{.05}=16.92$

For $v=9, \chi^2_{.01}=21.67$

Since, the calculated value of χ^2 is less than the table value at 5% and 1% level of significance, we accept null hypothesis and conclude that the variance of the distribution of weights of all students in the population is equal to 20 kgs.

Example 5: **A random sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5. Use $\alpha=0.05$ level of significance.**

Solution. We are given : $n=10, \Sigma(X-\bar{X})^2=50, \sigma^2=5$

Let us take the null hypothesis that the population variance is 5, *i.e.*

$H_0 : \sigma^2=5$ and $H_1 : \sigma^2>5$ ($\Rightarrow$ Right Tailed Test)

The χ^2- value is calculated as :

$$\chi^2=\frac{\Sigma(X-\bar{X})^2}{\sigma^2}$$

Putting the vlues, we have

$$\chi^2=\frac{50}{5}=10$$

Degrees of freedom $=v=n-1=10-1=9$

For $v=9, \chi^2_{.05}=16\cdot92$

Since, the calculated value of χ^2 is less than table value, we accept the null hypothesis and conclude that the variance of the population is 5.

EXERCISE – 1

1. A normal population has a standard deviation $\sigma=2.50$. A random sample of 12 values selected from this population yields sample variance $s^2=5.60$. Computer the X^2-value.

   [Ans. $\chi^2=10\cdot75$]

2. A sample of size $n=17$ drawn from a normally distributed population shows a variance $s^2=25$. Test the hypothesis that the population variance $\sigma^2=35$ against the alternative $\sigma^2>35$ using $\alpha=0.05$ level of significance.

   [Ans. $\chi^2=12\cdot14$, Accept H_0]

3. A random sample of size 10 from a normal population gives the following values :

   65, 72 68, 74, 77, 61, 63, 69, 73, 71

   Test the hypothesis that the population variance is 32 [Ans. $\chi^2=7\cdot316$, H_0 is accepted]

4. A sample of 20 observations gave a variance of 0.009. Is this compatible with the hypothesis that the sample is from a normal population with variance 0.010 ?

   [Ans.: $\chi^2=18$, Accept H_0]

5. A company producing TV tuners knows that on the average its product satisfactorily for 7 years, with a standard deviation of 1.75 years. A sample of 5 results in life times of 6, 8, 10, 7 and 9 years. Should the producer be satisfied that product still continues to have a standard deviations of 1.75 years ? accept the null

[Ans. $\chi^2 = 3{\cdot}26$, Accept H_0]

6. A random sample of size 25 from a population gives the sample S.D. to be 8.5 Test the hypothesis that the population S.D. is 10.

[Ans. $\chi^2 = 18{\cdot}06$, Accept H_0]

---

(2) χ^2-test as a non-paramtric test : χ^2-test is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom for using this test. As a non-parametric test, χ^2-test can be used (i) as a test of goodnesss of fit (ii) as a test of independence of attributes.

(i) χ^2-test as a test of goodness of fit : Under the test of goodness of fit, we try to find out how for the observed values of a given phenomenon are significantly different from the expected values *i.e.* there is good compatibility between theory and experiment or the fit is good. The term goodness of fit is also used for comparison of observed sample distribution with the expected probability distributions (such as Binomial, Poisson, Normal). χ^2-test determines how well theoretical distributions (such as Binomial, Poisson), fit the empirical distributions (*i.e.* those obtained from sample data)

Procedure :

(1) Set up the null hypothesis that there is no significant difference between the observed and the expected (or theoretical) values *i.e.* there is good compatibiliy between theory and experiment or the fit is good.

(2) We compute the value of χ^2 by using the formula :

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

Where, O = Observed frequency, E = Expected frequency

The above formula can also be written as :

$$\chi^2 = \Sigma\left[\frac{O^2}{E}\right] - N$$

Where, N is the total expected frequency and $\Sigma O = \Sigma E = N$

Note : The second form of the formula is more convenient for computation in case the expected frequencies comes in fractions.

(3) Degrees of freedom are worked out by using the following formula :

Degrees of freedon, $v = n - 1$

In case of Binomial, Poisson and Normal distributions, the degrees of freedom are obtained by subtracting the number of independent constraints from the total frequency (n). The number of independent constraints in a given data depends upon the number of parameters involved in the same data. This is indicated as under :

	Type of distribution	Constraints	No. of Constraints	Degrees of freedom
1.	Binomial distribution	Total frequency (n)	1	$n-1$
2.	Poisson distribution	Total frequency (n) and arithmetic mean (m)	2	$n-2$
3.	Normal distribution	n, $\bar{X}$ and σ	3	$n-3$

(4) The calculated value of χ^2 as such is than compared with the table value of χ^2 for give degrees of freedom at 5% and 1% level of significance. If the calculated χ^2 exceeds the table va of χ^2, we reject H_0 and conclude that the fit is not good. If the calculated value of χ^2 is less than t table value, we accept H_0 and conclude that the fit is good which means that the divergen between the observed and expected frequencies is attributable to fluctuations of sampling.

CONDITIONS FOR USING THE χ^2-TEST

χ^2-test as a goodness of fit can be used only when (i) n *i.e.* total frequency is large *i.e.* n > 50 The sample observations are independent (iii) The constraints on the cell frequencies, if any, a linear (iv) no theoretical (or expected) frequency should be small *i.e.* E < 5; if any E < 5, we us pooling technique *i.e.* we add the small frequencies with the preceding or succeeding frequency obtain the requried sum > 5 and adjust degrees of freedom (d.f.) accordingly.

Example 6 : **A die is thrown 180 times with the following results :**

No. turned up :	1	2	3	4	5	6	Total
Frequency :	25	35	40	22	32	26	180

Test the hypothesis that die is unbiased.

Solution. Set up the null hypothesis that the die is unbiased. On the basis of the hypothesis the expected frequency of each number turned up $= np = 180 \times \frac{1}{6} = 30$.

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
25	30	-5	25	0.833
35	30	+5	25	0.833
40	30	+10	100	3.333
22	30	-8	64	2.133
32	30	+2	4	0.133
26	30	-4	16	0.533
				$\Sigma(O-E)^2/E = 7.798$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 7.798$$

Degrees of freedom $= v = 6 - 1 = 5$

The tabulated value of χ^2 at 5% level of significance for 5 d.f. = 11.07

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that die is unbiased.

Example 7 : **The following figures show the distributions of digits in numbers chosen at random from a telephone directory :**

Digit :	0	1	2	3	4	5	6	7	8	9	Total
Frequency :	1026	1107	997	966	1075	933	1107	972	964	853	10,000

Test at 5% level whether the digits may be taken to occur equally frequently in the directory. (Given $\chi^2_{0.05}$ for 9 d.f. = 16.919).

Solution. Let us take the hypothesis that the digits may be taken to occur equally frequently in the directory. On the basis of this thypothesis, the expected frequencies are :

$$10{,}000 \times \frac{1}{10}, 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}$$

$$10{,}000 \times \frac{1}{10}, 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}$$

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2 / E$
1,026	1,000	+ 26	676	0·676
1,107	1,000	+ 107	11,449	11·449
997	1,000	-3	9	0009
966	1,000	- 36	1,156	1·156
1,075	1,000	+ 75	5,625	5·625
933	1,000	- 67	4,489	4·489
1,107	1,000	+ 107	11,449	11·449
972	1,000	- 28	784	0·784
964	1,000	- 36	1,296	0·296
853	1,000	- 147	21,609	21·609
				$\Sigma (O-E)^2 / E$ = 57·542

$$\therefore \quad \chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = 57 \cdot 542$$

Degrees of freedom $(v) = 10 - 1 = 9$

The table vlaue of χ^2 for 9 d.f. at 5% level of significance = 16·919.

Since, the calculated value of χ^2 is greater than the table value, we reject the hypothesis and conclude that the digits may not be taken to occur equally frequently in the directory.

Example 8 : **A sample analysis of examination results of 500 students were made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class. Are these figures commensurate with the**

general examination result which is in the ratio of 4 : 3 : 2 : 1 for the various categories respectively ?

(The table value of χ^2 for 3 d.f. at 5% level of significance is 7.81).

Solution. Let us take the hypothesis that the observed results are commensurate with the general examination results which is in the ratio of 4 : 3 : 2 : 1 .

The expected no. of students who have failed $= \frac{4}{10} \times 500 = 200$

The expected no. of students who have obtained a III class $= \frac{3}{10} \times 500 = 150$

The expected no. of students who have obtained a II class $= \frac{2}{10} \times 500 = 100$

The expected no. of students who have obtained a I class $= \frac{1}{10} \times 500 = 50$

Applying χ^2-test :

Category	O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
Failed	220	200	+ 20	400	2·000
3rd class	170	150	+ 20	400	2·667
2nd class	90	100	- 10	100	1·00
Ist class	20	50	- 30	900	18·00
					23·667

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 23 \cdot 667$$

Degrees of freedom$(v) = 4 - 1 = 3$

The tabulated value of χ^2 at 5% level of significance for 3 d.f. = 7·81.

Since, the calculated value of χ^2 is greater than the table value of χ^2, we reject the null hypothesis and conclude that the observed results are not commensurate with the general examination result.

Example 9 : **In an experiment on peas breeding, Mendal obtained the following frequencies of seeds : 315 round and yellow, 101 wrinked and yellow, 108 round and green, 32 wrinkled and green. According to his theory of heredity the numbers should be in proportion 9 : 3 : 3 : 1. Is there any evidence to doubt his theory at 5% level of significance ?**

Solution. Let us take the hypothesis that there is no siginficant difference in the observed and expected values. On the basis of this assumpation, the expected frequencies should be :

$$556 \times \frac{9}{16} = 312 \cdot 75,\ 556 \times \frac{3}{16} = 104 \cdot 25,\ 556 \times \frac{3}{16} = 104 \cdot 25,\ 556 \times \frac{1}{16} = 34 \cdot 75$$

Category	O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
Round and Yellow	315	312·75	2·25	5·0625	0·016
Wrinkled and Yellow	101	104·25	− 3·25	10·5625	0·101
Round and green	108	104·25	− 3·75	14·0625	0·135
Wrinkled and green	32	34·75	− 2·75	7·5625	0·218
	556				$\Sigma(O-E)^2/E$ = 0·47

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 0{\cdot}47$$

Degrees of freedom $= v = n - 1 = 4 - 1 = 3$

For $v = 3$, $\chi^2_{0.05} = 7.82$

Since, the calculated vlaue of χ^2 is less than the table value, we accept null hypothesis. Hence, there is no evidence to doubt the theory at 5% level of significance.

xample 10 : **The following table gives the number of aircraft accidents that occured during the various days of the week.**

Days of the week	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Total
No. of accidents	14	16	8	12	11	9	14	84

Find whether the accidents are uniformily distributed over the week:

(Given the table value of $\chi^2_{0.05}$ for 6 d.f. is 12·59)

olution. Let us take the hypothesis that the accidents are uniformily distributed over the week *i.e.* they are independent of the day of the week. On the basis of this hypothesis, we should expect 84/7 = 12 accidents on each day. Applying χ^2-test :

O	E	$(O-E)^2$	$(O-E)^2/E$
14	12	4	0·333
16	12	16	1·333
8	12	16	1·333
12	12	0	0·00
11	12	1	0·083
9	12	9	0·750
14	12	4	0·333
ΣO=84			$\Sigma(O-E)^2/E$ = 4·165

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 4{\cdot}165$$

Degrees of freedom $= v = n - 1 = 7 - 1 = 6$

For $v = 6$, $\chi^2_{0.05} = 12.59$

Since, the calculated value of χ^2 is less than the table value, we accept the n hypothesis and conclude that the accidents are uniformily distributed over week.

Example 11 : **The number of automobile accidents per week in a certain city were as follow**

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident's numbe were the same during these 10 week period.

Solution. Let us take the this hypothesis be that the number of accidents per week i certain are equal during the 10 week period.

On the basis of this hypothesis, the expected number of accidents per we

$$= \frac{100}{10} = 10.$$

Applying χ^2- test :

O	E	$(O-E)^2$	$(O-E)^2/E$
12	10	4	0·4
8	10	4	0·4
20	10	100	10·0
2	10	64	6·4
14	10	16	1·6
10	10	0	0·0
15	10	25	2·5
6	10	16	1·6
9	10	1	0·1
4	10	36	3·6
			$\Sigma(O-E)^2/E = 26{\cdot}6$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 26{\cdot}6$$

Degrees of freedom $= v = n - 1 = 10 - 1 = 9$

For $v = 9$, $\chi^2_{0.05} = 16{\cdot}92$

Since, the calculated value of χ^2 is greater than the table value, we reject the nu hypothesis and conclude that the accident conditions were not the same (uniform over the 10 week period.

Example 12 : **In a city 1000 children were born last week and out of these 600 were males and 400 females. Use chi-square test to assess the general hypothesis that the sex ratio for the newly born children is 1 : 1.**

Solution. Let us assume that the sex ratio for the newly born children is 1 : 1

Given : $N = 1000$ p = probability of a male child $= \frac{1}{2}$, q = probability of a female child $= \frac{1}{2}$. On the basis of the null hypothesis,

Expected no. of male child $= \frac{1}{2} \times 1000 = 500$

Expected no. of female child $= 1000 - 500 = 500$

	O	E	$(O-E)^2$	$(O-E)^2/E$
Males	600	500	10,000	20
Females	400	500	10,000	20
Total	1000	1000		$\chi^2 = 40$

Degrees of freedom $= v = n - 1 = 2 - 1 = 1$

For $x = 1$, $\chi^2_{.05} = 3.84$

Since, the calculated value of χ^2 is grater than the table value of χ^2, we reject the null hypothesis and conclude that the sex ratio for the newly bron children is not 1 : 1.

Test of Goodness of Fit of a Binomial Distribution :

Example 13 : **A set of 5 coins is tossed 3200 times and the number of heads appearing each time is noted. The result are given below :**

No. of heads	0	1	2	3	4	5
Frequency	80	570	1100	900	500	50

Solution. Let us take the null hypothesis that the coins are unbiased i.e. p (H) $= \frac{1}{2}$ On the basis of null hypothesis, the expected number of heads in a toss of 5 coins is calculated by the use of binomial distribution as follows :

$$P(x) = {}^nC_x \, q^{n-x} \cdot p^x \qquad \text{where } x = 0, 1, 2, 3, 4, 5$$

Given : $n = 5$, $N = 3200$, $p = p(H) = \frac{1}{2}$, $q = \frac{1}{2}$

X	$fe(x) = N \times {}^nC_x \cdot P(X)$	E
0	$3200 \times {}^5C_0 \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^0$	= 100
1	$3200 \times {}^5C_1 \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^1$	= 500
2	$3200 \times {}^5C_2 \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^2$	= 1000

3	$3200 \times {}^5C_3 \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^3$	= 1000
4	$3200 \times {}^5C_4 \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^4$	= 500
5	$3200 \times {}^5C_5 \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^5$	= 100

Applying χ^2- test :

O	E	$(O-E)^2$	$(O-E)^2/E$
80	100	400	4·00
570	500	4900	9·80
1100	1000	10,000	10·00
900	1000	10,000	10·00
500	500	0	0·00
50	100	2500	25·00
			$\Sigma\left[\frac{(O-E)^2}{E}\right] = 58·8$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 58·8$$

Degrees of freedom $= v = n - 1 = 6 - 1 = 5$

For $v = 5$, $\chi^2_{0.05} = 11·07$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that the coins are biased.

Example 13. **A random sample of 100 families with four children each disclosed the following data :**

No. of Female Births	0	1	2	3	4
No. of families	5	25	40	20	10

Verify at $\alpha = 0·05$ if these data are inconsistent with the hypothesis that male and female are equally likey.

Solution. Let us take the null hypothesis that the male and female births are equally probable *i.e.* $p = q = 1/2$ with 0, 1, 2, 3, 4 female basis.

On the basis of null hypothesis, the expected number of families can be calculated by the use of binomial distribution. The probability of x female birth in a family of 4 is given by :

$$P(x) = {}^nC_x \, q^{n-x} \, p^x \qquad \text{where, } x = 0, 1, 2, 3, 4,$$

Given : $n = 4$, $N = 100$, $p = \frac{1}{2}$, $q = \frac{1}{2}$

x	$fe(x) = N \cdot P(x)$	$E.$
0	$100 \times {}^4C_0 \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^0$	= 6·25

1	$100 \times {}^4C_1 \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^1$	= 25
2	$100 \times {}^4C_2 \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2$	= 37·5
3	$100 \times {}^5C_3 \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^3$	= 25
4	$100 \times {}^4C_4 \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^4$	= 6·25

Applying χ^2- test, we have

O	E	$(O-E)^2$	$(O-E)^2/E$
5	6·25	1·5625	0·25
25	25·0	0	0
40	37·5	6·25	0·17
20	25·0	25·0	1·00
10	6·25	14·0625	2·25
100			$\Sigma (O-E)^2/E = 3·67$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 3·67$$

Degrees of freedom $= v = n-1 = 5-1 = 4$

For $v=5$, $\chi^2_{0.05} = 9·49$

Since, the calculated value of χ^2 is less than the table value of χ^2, we accept the null hypotheis and conclude that the data are consistent with the hypothesis that male and fermale births are equally probable.

Test of Goodness of Fit of a Poisson Distribution :

Example 15 : **The number of defects per unit in a sample of 330 units of a manfactured product was found as follow:**

No. of defect :	0	1	2	3	4
No of units :	214	92	20	3	1

Fit a Poisson distribution to the data and test goodness of fit.

(Given $e^{-.439} = .6447$)

Solution.

Fitting of Poisson Distribution

x	f	fx
0	214	0
1	92	92
2	20	40
3	3	9
4	1	4
	$N = 330$	$\Sigma fx = 145$

$$\bar{X} = m = \frac{\Sigma fx}{N} = \frac{145}{330} = 0 \cdot 439$$

$\therefore$ Mean of the distribution $= m = 0 \cdot 439$

$$P(0) = e^{-m} = e^{-439} = \cdot 6447$$ [From the table]

By Poisson distribution, the expected frequencies are calculated as

$$fe(x) = N.P(x) = \frac{N.e^{-m} \cdot m^x}{x!}$$

Computation of Expected Frequencies

x	$fe(x) = N \times P(x)$	E
0	$f(0) = 330 \times e^{-439} = 330 \times 0 \cdot 6447$	$= 212 \cdot 75$
1	$f(1) = f(0) \cdot \frac{m}{1} = 212 \cdot 75 \times 0 \cdot 439$	$= 93 \cdot 4$
2	$f(2) = f(1) \cdot \frac{m}{2} = 93 \cdot 4 \times \frac{\cdot 439}{2}$	$= 20 \cdot 5$
3	$f(3) = f(2) \cdot \frac{m}{3} = 20 \cdot 5 \times \frac{\cdot 439}{3}$	$= 3 \cdot 00$
4	$f(4) = f(3) \cdot \frac{m}{4} = 3 \cdot 0 \times \frac{\cdot 439}{4}$	$= 0 \cdot 33$

After fitting Poisson distribution, we now apply χ^2 test of goodness of fit.

Let us take the null hypothesis that there is no significant difference between observed frequencies and the frequencies obtained by fitting Poisson distribution.

Appling χ^2-test, we have

Defects	O	E	$(O-E)^2$	$(O-E)^2 / E$
0	214	212·75	1·5625	0·0073
1	92	93·4	1·96	0·0210
2	20	20·5		
3	3 (24)	3·0 (23·83)	0·0289	0·0012
4	1	0·33		
				$\Sigma(O-E)^2/E = 0.0295$

In the above data, the frequencies of 3 and 4 defects are less than 5, so the frequencies for these defects have been pooled together with defects at 2 in order to make the sum total more than 5 or 5 and $E = 23 \cdot 83$.

Appling the formula.

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 0.0295$$

Degrees of freedom $= v = n - 2 = 3 - 2 = 1$

[Since, after grouping only 3 classes are left, therefore $n = 3$]

For $v = 1$, $\chi^2_{.05} = 3.84$

Since, the calculated value of χ^2 is much less than the table value, we accept the null hypothesis and conclude that the fit is good.

EXERCISE – 2

1. A die is thrown 100 times and the frequency of various faces are given as below :

Face	1	2	3	4	5	6
Frequency	17	14	20	17	17	15

Test the hypothesis that die is unbiased. Use 5% l.o.s. [Ans. $\chi^2 = 1.2796$, Accept H_0]

2. 200 digits were chosen at random from a set of tables. The frequencies of the digits were as follows :

Digits	0	1	2	3	4	5	6	7	8	9	Total
Frequencies	18	19	23	21	16	25	22	20	21	15	200

Using χ^2-test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the table from which they were drawn (The 5% value of χ^2 for 9 d.f. is 16·92).

[Ans. $\chi^2 = 4.30$, Accept H_0]

3. A research investigator selected a random sample of 200 voters to find which political party they would vote in the municipal elections. The results were observed as under :

Political Party	A	B	C	D
No. of voters	40	90	50	20

Verify at $\alpha = 0.05$ if the observed data provide sufficient evidence that the four political parties are equaly preferred. [Ans. $\chi^2 = 4.30$, Accept H_0]

4. In an experiment on peas breeding, the following frequencies of feeds were obtained : 218 round and yellow; 72 wrinkled and yellow; 90 round and green, 20 wrinkled and green, Total 400. Theory products that the frequencies should be in the proportions 9 : 3 : 3 : 1. Examine the difference between theory and experiment (Value of χ^2 at 4 and 3 degrees of freedom are 9·448 and 7·815) at 5% level of significance. [Ans. $\chi^2 = 4.338$, Accept H_0]

5. The theory predicts the proportion of beans in the four groups, A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the number in the four groups were 882, 313, 287 and 118. Does the experiment result support the theory ? Apply χ^2-test.

[$\chi^2 = 4.7226$, Accept H_0]

6. The following tables gives the number of books borrowed from a public library during a particular week :

Days of the week	Mon	Tue	Wed	Thu	Fri	Sat
No. of Books borrowed	140	132	160	148	134	150

Test the hypothesis that the number of books borrowed does not depend on the day of the week. Test at 5% level of significance. [Ans. $\chi^2 = 3{\cdot}941$, Accept H_0]

7. A survey of 320 families with 5 children each revealed the following distribution :

No. of boys :	5	4	3	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable ? [Ans. $\chi^2 = 7{\cdot}16$. Accept H_0]

8. 4 coins were tossed 160 times and the following results were obtained :

No. of heads :	0	1	2	3	4
No. of obseved frequency :	17	52	54	31	6

Under the assumption that coins are unbiased, find the expected frequencies of getting 0, 1, 2, 3 or 4 heads and test the goodness of fit. [Ans. $\chi^2 = 12{\cdot}725$, Reject H_0]

9. A book has 700 pages. The number of pages with various number of misprints is recorded below. At 5% significance level, are the misprints distributed according to possion law :

No. of misprints :	0	1	2	3	4	5	Total
No. of pages with x misprints :	616	70	10	2	1	1	700

[Ans. $\chi^2 = 11{\cdot}04$, Reject H_0, Fit is not good]

10. The following mistakes per page were observed in a book :

No. of mistakes per page :	0	1	2	3	4	Total
No. of units :	211	90	19	5	0	325

Does this information verify that the mistakes are distributed according to Poisson distribution ? [$e^{-0{\cdot}44} = 0{\cdot}644$]

[Ans. $\chi^2 = 0{\cdot}07$, Accept H_0]

11. The number of car accidents in a metropolitan city was found as 20, 17, 12, 6, 7, 15, 8, 5, 16 and 14 per month respectively. Use chi square test to check whether these frequencies are in agreement with the belief that occurrence of accidents was the same during the 10 months period. Test at 5% level of significance. (Take value at 5% level for $v = 9$ is 16.9)

[Ans. $\chi^2 = 20{\cdot}331$ Reject H_0]

12. The following grades were given to a class of 100 students

Grade :	A	B	C	D	E
Frequency :	14	18	32	20	16

Test the hypothesis at the 0·05 level, that the distribution of grade is uniform.

Given :

Degree of Freedom :	3	4	5
χ^2-value :	7·81	9·49	11·07

[Ans. $\chi^2 = 10$, Raject H_0]

13. In the accounting department of a bank 100 accounts are selected at random and examined for errors. Suppose the following results have been obtained:

No. of errors :	0	1	2	3	4	5	6
No. of accounts :	36	40	19	2	0	2	1

On the basis of this information can it be concluded that the errors are distributed according to the poisson probability law ? [Ans. $\chi^2 = 1·450$, Accept H_0]

14. A survey of 200 families with 3 children selected at random gave the following results.

Male Births	0	1	2	3
No. of families	40	58	62	40

Test the hypothesis that male and female are equally likely at 5% level of significance.
[Ans. $\chi^2 = 24·1$, Reject H_0]

15. In a city the percentage of smokers was 90. A random sample of 100 persons was taken and out of them universe 85 were found smokers. Use chi square test and tell whether sample ratio significantly differs from the universe ratio for the city [Ans. $\chi^2 = 2.778$, Accept H_0]

16. The manager of a theatre complex with four theatres wanted to see whether there was a difference in popularity of the four movies currently showing for saturday afternoon matinees. The number of custmers for each movie was recorded for one Satuday afternoon with the following results : 63, 55, 75 and 77 customers viewed movies, 1, 2, 3 and 4 respectively. Complete the test to see whether there is a difference at the 5% level of significance. [Ans. $\chi^2 = 4.78$, Reject H_0]

(ii) χ^2-test as a test of independence of attributes : χ^2-test enables us to examine whether or not two attributes are associated or independent of one another. For example, we may be interested in knowing whether a new medicine is effective in controlling fever or not. χ^2-test will help us in deciding this issue.

Procedure :

(i) Set up the null hypothesis that the two attributes (*viz.* new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever.

(ii) On the basis of the null hypothesis, we calculate the expected frequencies by using the following formula :

$$\text{Expected frequency} = \frac{(R) \times (C)}{N}$$

Where, R = Row Total, C = Column total, N = Total number of observations.

(iii) We compute the χ^2-value by using the following formula :

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

(iv) For a contingency table which has 'r' rows and 'c' columns, degrees of freedom are worked out by using the formula.
Degrees of freedom $= v = (c-1)(r-1)$

(v) Obtain the critical value (or table value) of χ^2 with reference to the degrees of freedom for the given problem and the desired level of significance.

(vi) If the calculated value of χ^2 is less than the table value at a certain level of significance for given degrees of freedom, we accept the null hypothesis and conclude that the two attributes are independent (*i.e.* the new medicine is not effective in controlling the fever). But if the calculated value χ^2 is greater than its table value, we reject the null hypothesis and conclude that the two attributes are associated (*i.e.* new medicine is effective in controlling fever).

Example 16 : **A sample of 200 persons with a particular disease was selected. Out of them, 100 were given drug and others were not. The results were observed as follows:**

	No. of Persons Given		
	Drug	No Drug	Total
Curred	55	65	120
Not Cured	45	35	80
Total	100	100	200

Test whether the drug has been effective in curing the disease.

Solution. Let the null hypothesis be that drug has not been effective in curing the disease. On the basis of this hypothesis, the expected frequencies are calculated as follow:

$$E_{11} = \frac{100 \times 120}{200} = 60$$

The remaining frequencies can be found by subtractions from the column and row totals.

The expected frequencies table would be as follows :

60	60	120
40	40	80
100	100	200

Applying χ^2- test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
55	60	− 5	25	0·416
45	40	+ 5	25	0·625
65	60	+ 5	25	0·416
35	40	− 5	25	0·625
				$\Sigma (O-E)^2/E = 2·082$

$\therefore$

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2·082$$

Degrees of freedom $= v = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3 \cdot 84$

Since, the calculated value of χ^2 is less than the table value of χ^2, we accept the null hypothesis and conclude that the drug has not been effective in curing the disease.

Example 17 : The table given below shows the data during and epidemic of cholera :

	Attacked	Not Attacked	Total
Inoculated	31	469	500
Not Inoculated	185	1,315	1,500
Total	216	1,784	2,000

Use χ^2 test to determine whether inoculation is effective in preventing the attack of cholera (Given as 5% level of significance, the value of $\chi^2_{0.05}$ for 1 d.f. = 3·84).

Solution. Let us take the hypothesis that inoculation is not effective in preventing the attack of cholera. On the basis of this hypothesis, the expected frequencies are :

$$E_{11} = \frac{216 \times 500}{2,000} = 54$$

The remaining frequencies can be found out by subtractions from the column and row totals.

The expected frequecies table would be as follows:

54	446	500
162	1338	1500
216	1784	2000

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
31	54	–23	529	9·797
185	162	+ 23	529	3·266
469	446	+ 23	529	1·187
1315	1338	– 23	529	0·396
				$\Sigma (O-E)^2/E = 14 \cdot 646$

$$\therefore \quad \chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = 14 \cdot 464$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3 \cdot 84$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that inoculation is effective in preventing the attack of cholera.

Example 18 : Two investigators study the income of group of persons by the method of sampli Following results were obtained by them :

Investigator	Poor	Middle class	Well-to-do	Total
A	160	30	10	200
B	140	120	40	300
Total	300	150	50	500

Show that the sampling technique of at least one of the investigator suspected.

(Given the value of $\chi^2_{0.05}$ for 2 d.f. = 5·991)

Solution. Let us take the hypothesis that there is no suspicion about the sampling techniq of the two investigators. On the basis of this hypothesis, the expected frequenc shall be :

$$E_{11} = \frac{300 \times 200}{500} = 120, \qquad E_{12} = \frac{150 \times 200}{500} = 60$$

The remaining frequencies can be found out by subtractions from the column a row totals.

The table of expected frequencies is given below :

120	60	20	200
180	90	30	300
300	150	50	500

applying χ^2-test

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
160	120	40	1600	13·333
140	180	− 40	1600	8·888
30	60	− 30	900	15·000
120	90	30	900	10·000
10	20	−10	100	5·000
40	30	+ 10	100	3·333
				$\Sigma (O-E)^2/E = 55·554$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 55·554$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(3-1) = 2$

For $v = 2$, $\chi^2_{0.05} = 5·991$

Since, the calculated value of χ^2 is greater than the table value, we reject the n hypothesis and conclude that the sampling technique of at least one of th investigators is suspected.

Example 19 : A milk producer's union wishes to test whether the preference pattern of consumers for its product is dependent on income levels. A random sample of 500 individuals gives the following data :

Income	Product Preferred			
	Product A	Product B	Product C	Total
Low	170	30	80	280
Medium	50	25	60	135
High	20	10	55	85
Total	240	65	195	500

Can you conclude that the preference patterns are independent of income levels?

(For $v = 4$, $\chi^2_{0.05} = 9.49$)

Solution. Let us take the hypothesis that preference patterns are independent of income levels. On the basis of this hypothesis, the expected frequencies corresponding to different rows and columns shall be :

$$E_{11} = \frac{240 \times 280}{500} = 134.4 \qquad E_{12} = \frac{65 \times 280}{500} = 36.4$$

$$E_{21} = \frac{240 \times 135}{500} = 64.8 \qquad E_{22} = \frac{65 \times 135}{500} = 17.55$$

134.40	36.40	109.20	280
64.80	17.55	52.65	135
40.80	11.05	33.15	85
240	65.00	195.00	500

Applying χ^2- test :

O	E	$(O-E)^2$	$(O-E)^2/E$
170	134.40	1267.36	9.430
50	64.80	219.04	3.3802
20	40.80	432.64	10.603
30	36.40	40.96	1.125
25	17.55	55.50	3.162
10	11.05	1.10	0.099
80	109.20	852.64	7.808
60	52.65	54.02	1.026
55	33.15	477.42	14.402
			$\Sigma (O-E)^2/E = 51.036$

$$\therefore \quad \chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = 51.036$$

Degrees of freedom $= v = (r-1)(c-1) = (3-1)(3-1) = 4$

For $v = 4$, $\chi^2_{0.05} = 14 \cdot 860$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and hence conclude that preference patterns are not independent of income levels.

Example 20 : **In a survey of 200 boys, of which 75 were intelligent, 40 had educated fathers; while 85 of the unintelligent boys had uneducated fathers. Do these figures support the hypothesis that educated fathers have intelligent boys. Use χ^2-test (The value of χ^2 for 1 degree of freedom at 5% level is 3 · 84).**

Solution. The given data can be tabulated as follows :

Boys/Fathers	Educated	Uneducated	Total
Intelligent	40	35	75
Unintelligent	40	85	125
Total	80	120	200

Let us take the hypothesis that there is no association between the education of fathers and intelligence of sons.

On the basis of this hypothesis, the expected frequencies shall be :

$$E_{11} = \frac{75 \times 80}{200} = 30$$

The remaining frequencies can be found be subtracting from the column and row totals.

The table of expected frequencies shall be as follows :

30	45	75
50	75	125
80	120	200

Applying χ^2-test :

O	E	$(O-E)^2$	$(O-E)^2/E$
40	30	100	3 · 333
40	50	100	2 · 000
35	45	100	2 · 222
85	75	100	1 · 333
			$\Sigma (O-E)^2/E = 8 \cdot 888$

$\therefore$

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = 8 \cdot 888$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3 \cdot 84$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and hence that educated fathers have intelligent boys.

Alternative Formula for Finding the Value of χ^2 in a (2×2) table

There is an alternative formula of calculating the value of χ^2 in a (2×2) table. If we write the cell frequencies and marginal totals in case of a (2×2) table as :

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

then the formula for calculating the value of χ^2 will be written as follows :

$$\chi^2 = \frac{N.(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \qquad \text{where, } N = a+b+c+d$$

Note : The alternative formula is rarely used in finding the value of χ^2 as it is not applicable uniformaly in all cases but can be used only in a 2×2 contingency table.

Example 21. **In an anti-malaria campaign in a certain area, quinine was administrated to 812 persons out of a total population of 3248. The number of fever cases is shown below :**

Treatment	Fever	No Fever	Total
Quinine	20	792	812
No quinine	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of quinine in checking malaria. (Given for $v=1$, $\chi^2_{0.05} = 3.84$)

Solution. Let us take the null hypothesis that the quinine is not effetive in checking malaria. Arrange the given data in a designated form, we have

	Fever	No Fever	Total
Quinine	a 20	b 792	$a+b$ 812
No Quinine	c 220	d 2216	$c+d$ 2436
Total	$a+c$ 240	$b+d$ 3008	N 3248

For 2×2 table, using the direct formula of computing χ^2, we have

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Putting the values, we have

$$= \frac{3248(20 \times 2216 - 220 \times 792)^2}{(240)(3008)(812)(2436)} = 38.48$$

Degrees of freedom $= v = (c-1)(r-1) = (2-1)(2-1) = 1$

For $v=1$, $\chi^2_{0.05} = 3.84$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that the quinine is usefulness in checking malaria.

YATES CORRECTIONS IN A 2×2 TABLE

F. Yates has suggested corrections for continuity in χ^2 value calculated in a 2×2 table, particularly, when any cell frequency is less than 5. The correction suggested by Yates is popularly known as **Yates' Correction**. It involves the reduction of the deviation of observed from expected frequencies which of course reduces the value of χ^2. The rule for correction is to increase the observed frequencies which is less than 5 by 0·5 and than the remaining frequencies are adjusted by adding or subtracting 0·5 to them without disturbing the marginal totals. The observed values, thus corrected will be represented by O from which deviations of the corresponding expected values, E will be found.

Note : In a 2×2 table, the method of pooling cannot be applied.

Example 22: **The result of a certain survey of 50 ordinary shops of small size is given below:**

	Shops in		Total
	Towns	Villages	
Run by Men	17	18	35
Run by Women	3	12	15
Total	20	30	50

Can it be said that shops run by women are relatively more in villages than in towns. Use χ^2-test. (Table value of $\chi^2_{0.05}$ for one degree of freedom at 5% level of significance is 3·84).

Soluiton : Let us take the null hypothesis that shops run by women are equal in number in villages as well as in towns. On the basis of this hypothesis, the expected frequencies will be as follows :

$$E_{11} = \frac{20 \times 35}{50} = 14$$

The remaining frequencies can be found out by subtractions from the column and row totals.

The table of expected freqencies will be :

14	21	35
6	9	15
20	30	50

Since, one of the observed frequency is less 5, we increase the value of that observed frequency by 0·5 and adjust other frequencies using Yates' corrections. The adjusted observed frequencies after Yates' corrections will be as follow :

17 − 0·5 = 16·5	18 + 0·5 = 18·5	35
3 + 0·5 = 3·5	12 − 0·5 = 11·5	15
20	30	50

With the above expected and corrected observed values, the corrected value of χ^2 will be obtained as :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
16·5	14	2·5	6·25	0·446
3·5	6	− 2·5	6·25	1·042
18·5	21	− 2·5	6·25	0·298
11·5	9	+ 2·5	6·25	0·694
				$\Sigma(O-E)^2/E = 2·48$

$$\therefore \quad \chi^2 = \Sigma\frac{(O-E)^2}{E} = 2·48$$

Degrees of freedom $= v = (c-1)(r-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3·84$

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that number of shops run by women are not relativiely more in villages than in towns.

Example 23 : **In an experiment on the immunization of goats from Anthrax, the following results were obtained. Derive your inference on the efficacy of the vaccine :**

	Diet of Anthrax	Survived	Total
Inoculation with vaccine	**2**	**10**	**12**
Not inoculated	**6**	**6**	**12**
	8	**16**	**24**

Solution. It is quite obvious from the above data that Yates' correction shall be applied here. Let us take the null hypothisis that there is no relationship between Inoculation with vaccine and death from anthrax.

Observed frequencies

2	10	12
6	6	12
8	16	24

Observed frequencies with Yates' corrections

2.5	9.5	12
5.5	6.5	12
8	16	24

Expection frequencies

$\frac{12\times 8}{24} = 4$	12 − 4 = 8	12
8 − 4 = 4	12 − 4 = 8	12
8	16	24

$$\chi^2 = \frac{(2.5-4)^2}{4} + \frac{(9.5-8)^2}{8} + \frac{(5.5-4)^2}{4} + \frac{(6.5-8)^2}{8}$$

$$= 0.56250 + 0.28125 + 0.56250 + 0.28125 = 1.6875$$

Degrees of freedom $= v = (c-1)(r-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{.05} = 3.84$

Since the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that thre is no relationship between inoculation with vaccine and death from anthrax *i.e.* immunization is not effective.

Test of Equality of Several Population Proportions

Testing of equality of two or more population poportions is an extension of the χ^2-test of independence. χ^2-test is also used to examine the equality two or more population proportions. The following examples clarify the procedure.

Example 24 : **A social organisation claiming to be the promoters of sex education sought the views of parents from the states of Punjab, Bihar and Haryana introducing sex education at the school level. The views of 80 parents selected at random from each of the three states are as follows :**

	Punjab	Bihar	Haryana
In favour	50	20	45
Against	30	60	35

Do the sample provide enough evidence to the view that the proportion of parents in favour of introducing sex education in schools is the same in all three states ? Use $\alpha = \cdot 01$.

Solution. Let the null hypothesis be that the proportion of parents in favour of sex education in schools is the same in the three states.

One the basis of this hypothesis, the expected frequencies are calculated as follows:

$$E_{11} = \frac{80 \times 115}{240} = 38.3 \qquad E_{12} = \frac{80 \times 115}{240} = 38.3$$

The remaining frequencies can be found out by subtracting from the column and row totals.

The expected frequencies would be as follows :

38·3	38·3	38·4	115
41·7	41·7	41·6	125
80	80	80	240

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
50	38.3	11.7	136.89	3.57
20	38.3	−18.3	334.89	8.74
45	38.4	6.7	44.89	1.169
30	41.7	−11.7	136.89	3.28
60	41.7	18.3	334.89	8.03
35	41.6	− 6.6	43.56	1.04
				$\Sigma[(O-E)^2/E] = 25.83$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 25 \cdot 83$$

Degrees of freedom $= v = (c-1)(r-1) = (3-1)(2-1) = 2$

For $v = 2$, $\chi^2_{0.01} 0 \cdot 1 = 9 \cdot 21$

Since, the calculated value of χ^2 is less than the table value of χ^2, we reject H_0 and conclude that the proportions of parents in favour of introducing education at the school level are not the same in the three states.

Example 25. **The following data gives the HDL-level in random samples of sizes 120, 200, 150 and 130 from the adult population of the four cities A, B, C and D.**

	A	B	C	D
High HDL	53	80	68	57
Not High DHL	67	120	82	73

Test the equality of proportions of adults with high HDL Cholesterol in these four cities. Use $\alpha = 0.025$.

Solution. Let P_1, P_2, P_3 and P_4 represents the true proportions of adults with high HDL cholesterol in the cities A, B, C and D respectively.

Set up the hypothesis :

Null hypothesis : $H_0 : P_1 = P_2 = P_3 = P_4$

Alternative hypothesis : $H_1 : P_1, P_2, P_3$ and P_4 are not all equal.

Compute the expected frequency for each observed frequency by the formula (under the hypothesis of independence) :

$$\text{Expected Frequency} = \frac{(\text{Row Total})\ (\text{Columm Total})}{\text{Grand Total}}$$

The observed and the expected frequencies are given in Table below. The bold figures in brackets ( ), represent the corresponding expected frequencies.

Observed and Expected Frequencies

	A	B	C	D	Total
High HDL	53 (51.6)	80 (86)	68 (64.5)	57 (55.9)	258
Not High HDL	67 (68.4)	120 (144)	82 (85.5)	73 (74.1)	342
Total	120	200	150	130	600

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$= \frac{(53-51.6)^2}{51.6} + \frac{(80-86)^2}{86} + \frac{(68-64.5)^2}{64.5} + \frac{(57-55.9)^2}{55.9}$$

$$+ \frac{(67-68.4)^2}{68.4} + \frac{(120-114)^2}{114} + \frac{(82-85.5)^2}{85.5} + \frac{(73-74.1)^2}{74.1}$$

$$= 0.0380 + 0.4186 + 0.1899 + 0.0216 + 0.0287 + 0.3158 + 0.1433 + 0.0163 = 1.1722$$

Degrees of freedom : $v = (r-1)\ (c-1) = (2-1)\ (4-1) = 3$

The critical value of chi square for 3 d.f. and level of significance 0.025 is $\chi^2_{3(0.025)} = 9.348$.

Since, the computed value of test statistic is less than the critical (tabulated) valu it is not significant. Hence, we fail to reject the null hypothesis at 0.025 level o significance.

Conclusion : H_0 may be accepted at level of significance $\alpha = 0.025$ and we ma conclude that the proportion of adults with high HDL cholesterol level is mos likely the same in all the four cities.

Example 26. **It is found that 35 of 250 housewives in Delhi, 22 of 220 housewives in Mumba and 39 of 300 housewives in Chandigarh watch at least one talk show every day At the 0.05 level of significance, test that there is no difference between the true proportions of housewives who watch talk shows in these cities.**

Solution. Let P_1, P_2 and P_3 represent the true proportion of housewives who watch talk shows in the cities of Delhi, Mumbai and Chandigarh, respectively.

Null hypothesis : $H_0 : P_1 = P_2 = P_3$

Alternative hypothesis : $H_1 : P_1, P_2$ and P_3 are not all equal.

Expected frequencies : Compute the expected frequency for each of the cell frequencies by the formula (under the hypothesis of independence) :

$$\text{Expected Frequency} = \frac{\text{(Row Total) (Columm Total)}}{\text{Grand Total}}$$

The observed frequencies, along with the expected frquencies [(in the bold in brackets ()] are given in the following Table.

Observed and Expected Frequencies

	Delhi	Mumbai	Chandigarh	Total
Watch Talk Show	35 **(31.2)**	22 **(27.4)**	39 **(37.4)**	96
Do not Watch Talk Show	215 **(218.8)**	198 **(192.6)**	261 **(262.6)**	674
Total	250	220	300	770

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$= \frac{(35-31.2)^2}{31.2} + \frac{(22-27.4)^2}{27.4} + \frac{(39-37.4)^2}{37.4} + \frac{(215-218.8)^2}{218.8} + \frac{(198-192.6)^2}{192.6} + \frac{(261-262.6)^2}{262.6}$$

$= 0.4628 + 1.0642 + 0.0684 + 0.0660 + 0.1514 + 0.0097 = 1.8225.$

Degrees of freedom : $v = (r-1)(c-1) = (2-1)(3-1) = 2$

The critical value of chi-square for 2 *d.f.* and 0.05 level of significance is 5.991.

Conclusion : Since, the calcluated value of test statistic ($\chi^2 = 1.8225$ is less than the tabulated value) $\chi^2 = 5.991$, it is not significant. Thus, the data do not provide enough evidence against the null hypothesis, which may be accepted at 5% level

of significance. Hence, we conclude that the proportion of housewives who watch talk shows is same in all the three cities.

TEST OF HOMOGENITY

The test of homogenity is another extension of the χ^2-test of independence. Tests of homogenity are used to determine whether two or more independent random samples are drawn from the sample population. Instead of one sample as we use with independence problem, we shall now have two or more samples are from each population.

The following example clarify the procedure of the test :

Example 27 : **A insurance company has introduced a new scheme for employees. Independent random samples of 100 males and 120 females when examined to know their views about the new scheme yielded the following results :**

	For	Against	Indifferent	Total
Male	25	40	35	100
Female	35	55	30	120
Total	60	95	65	220

Test the hypothesis at $\alpha = \cdot 01$ that the two samples have come from a homogenous populations.

Solution. Let us take the null hypothesis that the two samples have come from a homogenous population. On the basis of the hypothesis, The expected frequencies are calculated as :

$$E_{11} = \frac{60 \times 100}{220} = 27.3 \qquad E_{12} = \frac{95 \times 100}{220} = 43.2$$

The remaining frequencies can be found out by subtracting from the column and row totals.

The expected frequencies worked be as follows :

27·3	43·2	29·5	100
32·7	51·8	35·5	120
60	95	65	220

Applying χ^2-test :

O	E	(O – E)	$(O - E)^2$	$(O - E)^2 / E$
25	27·3	– 2·3	5·29	0·1937
40	43·2	– 3·2	10·24	0·2370
35	29·5	5·5	30·25	1·025
35	32·8	2·3	5·29	0·1617
55	51·8	3·2	10·24	0·1976
30	35·5	5·5	30·25	0·8521
				$(O - E)^2 / E = 2·6671$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2{\cdot}6671$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{.05} = 3{\cdot}84$

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that the two samples have come from homogenous populations.

EXERCISE – 3

1. A survey amongst women was conducted to study the family life. The observations are as follows :

Education	Family Life		
	Happy	Not Happy	Total
Educated	70	30	100
Non-Educated	60	40	100
Total	130	70	200

Test whether there is any association between familiy life and education.
(The table value of $\chi^2_{0{\cdot}05}$ for 1 d.f. = 3·84)

[Ans. $\chi^2 = 2{\cdot}198$, Accept H_0]

2. Calculate the expected frquencies for the following data persuming the two attributes, *viz*, condition of home and condition of child as independent.

		Condition of Home	
		Clean	Dirty
Condition of Child	Clean	70	50
	Fairly Clean	80	20
	Dirty	35	45

Use chi-square test at 5% level of signifiance whether the two attributes are indendent.
(Table values of chi-square at 5% for 2 d.f. is 5.991 and for 3 d.f. is 7.815 and for 4 d.f. is 9.488.)

[Ans. $\chi^2 = 25{\cdot}848$, Rejeet H_0]

3. Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence level. The results are as follows :

Researcher	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	137	164	152	147	600
Y	32	57	56	35	180
Total	169	221	208	182	780

Would you say that the sampling techniques adopted by the two researchers are significantly different? (For $v = 3, \chi^2_{0.05} = 7.82$ [Ans. $\chi^2 = 5.801$, not different]

4. From the following data, find out whether there is any relationship between sex and prefenece for colour :

Colour	Males	Females	Total
Green	40	60	100
White	35	25	60
Yellow	25	15	40
Total	100	100	200

(Given for $v = 2, \chi^2_{0.05} = 5.991$) [Ans. $\chi^2 = 8.166$, Reject H_0]

5. A sample of 400 students of under-graduate and 400 staudents of post-graduate classes was taken to know their opinion about autonomos colleges. 290 of the under-graduate and 310 of the post-graduate students favoured the autonomous status. Present these facts in the form of a table and test, at 5% level, that the opinion regarding autonomous status of colleges is independent of the level of classes of students. (Table value χ^2 at 5% level is 3.84 for 1 *d.f.*). [Ans. $\chi^2 = 2.66$, Accept H_0]

6. Two treatments A and B were tried to control a certain type of plant disease. The following results were obtained.

A : 400 plants were examined and 80 were found infected.

B : 400 plants were examined and 70 were found infected.

Is the treatment B superior to treatment A ?

(Given that $\chi^2_{0.05}(1) = 3.84; \chi^2_{0.05}(3) = 7.82$) [Ans. $\chi^2 = 0.82$, Accept H_0]

7. In an experiment on immunization of cattle from tuberculosis, the following were obtained :

	Affected	Not Affected	Total
Inoculated	4	20	24
No inoculated	6	50	56
Total	10	70	80

Calculate χ^2 and discuss the effect of vaccine in controlling suspectability to tuberculosis. [Applying Yates' correction] [Ans. $\chi^2 = 2.04$, Accept H_0]

8. The following table gives the frequencies of firms on automation and productivity :

	Productivity increased	Productivity not increased	Total
Automated	32	468	500
Not Automated	184	1316	1500
Total	216	1784	2000

Use χ^2 (Chi-Square) test to determine whether productivity is independent of the automation ($\chi^2_{0.05}$ at 1 d.f. = 3.84) [Ans. $\chi^2 = 13.395$, Reject H_0]

9. A drug is said to be useful for the treatment of cold. In an experiment carried out on 160 persons suffering from cold, half of the persons were treated with the drug and rest half with sugar pills. The effect of treatment is described in the following table :

	Helped	Harmful	No Effect
Drug	52	10	18
Suger Pills	44	10	26

Test the hypothesis that in the treatment of cold the drug is not at all effective as compared to sugar pills.

[Given for $v = 2$, $\therefore \chi^2_{.05} \cdot 05 = 5 \cdot 991$] [Ans. $\chi^2 = 2 \cdot 12$, Accept H_0]

10. Two sample polls of votes for two candidates A and B for public office are taken, one each from among residents of rural and urban areas. The results are given below. Examine whether the nature of area is related to voting preference in this election ?

Area/Candidate	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Given $\chi^2_{0\,05} = 3 \cdot 841, 5 \cdot 991, 7 \cdot 82$ for 1, 2 and 3 d.f. respectively.

[Ans. $\chi^2 = 10 \cdot 32$, Accept H_0]

11. Two groups of 100 people each were selected for testing the use of a vaccine. 15 persons contracted the disease out of the inoculated persons in one group, while 25 contracted the disease in the other group. Test the efficacy of the vaccine using the χ^2 test.

(Given of $v = 1$, $\chi^2_{0\,05} = 3 \cdot 84$) [Ans. $\chi^2 = 3 \cdot 124$, Accept H_0]

12. A company has head offices at three places : A, B and C, A random sample of 10 executives posted at A, of 12 posted at B, and of 15 posted at C were examined to find out the number of those suffering from hypertension. The sample results were found to be as given below:

	A	B	C
Hypertension cases	4	7	9
Non-hpertension cases	6	5	6

Verify at $\alpha = 0 \cdot 01$ if the proportion of executives suffering from the hypertension at three head offices are the same. [Ans. $\chi^2 = 1 \cdot 0975$, Accept H_0]

13. From the adult male population of four large cities, random samples of sizes given below were taken and the number of married and single men recorded. Can we say that the proportion of married men is same in all the four cities?

City →	A	B	C	D	Total
Married	137	164	152	147	600
Single	32	57	56	35	180
Total	169	221	208	182	780

[Ans. $\chi^2 = 5.801$, Accept H_0]

ADDITIVE PROPERTY OF χ^2

Chi-Square possesses the additive property. If a number of samples of similar data have been independently collected and a number of χ^2 values have been obtained there form, it is possible to combine them by the simple process of addition. This helps in getting a better idea about the significance or otherwise of the problem in hand as instead of on investigation (or one sample).

Example 28. **An investigation was made in eight big cities of a state with a view to test the effectiveness of inoculation during an epidmic of cholera. The following results were obtained :**

Cities	A	B	C	D	E	F	G	H
χ^2 value	2·32	3·64	3·15	4·54	2·24	3·66	4·87	6·72
d.f.	1	1	1	1	1	1	1	1

Find out the pooled χ^2 for all the eight cities of any state and test your result at 5% level of significance.

Solution. $H_0 : f_0 = f_e$ (Observed and expected distributions are the same)

$H_1 : f_0 \neq f_e$ (Difference between observed and expected distributions is significant)

$\alpha = \cdot 05$; d.f. $= 8, \chi^2 = 15 \cdot 507$

Cities	A	B	C	D	E	F	G	H	Total
χ^2 value	2·32	3·64	3·15	4·54	2·24	3·66	4·87	6·72	Pooled χ^2 = 31·14
d.f.	1	1	1	1	1	1	1	1	8

INTERPRETATION

The table value of χ^2 at 5% level of significance with 1 d.f. is 3·841 and 8 d.f. is 15·507. By the analysis of each city separately, it is clear that the difference is not significant in the cities A, B, C, E and F *i.e.* the null hypothesis is true where as the difference in cities D, G and H is significant and the null hypothesis is not ture.

But the combined (or pooled) calculated χ^2 is 31·14 which is greater than 15·507. Thus combined value is greater than the table value; hence the difference in the cities together is significant. That is the null hypothesis is not true by considering all the cities together.

MISCELLANCEOUS SOLVED EXAMPLES

Example 29 : **A controlled experiment was conducted to test the effectivencess of a new drug. Under this experiment 300 patients were treated with the new drug and 200 were not treated with the drug. The results of the experiments are presented below. Using the Chi-square test, comment on the effectiveness of drug.**

Details	Cured	Condition worsened	No effect	Total
Treated with the new drug	200	40	60	300
Not treated with the new drug	120	30	50	200
Total	320	70	110	500

Solution. Let us take the hypothesis that the new drug is not effective. On the basis of this hypothesis, the expected frequencies are calculated as follows :

$$E_{11} = \frac{320 \times 300}{500} = 192; \qquad E_{12} = \frac{70 \times 300}{500} = 42 \text{ and so on.}$$

The remaining frequencies can be found out by subtraction from the column and row totals.

The expected frequencies is given below :

192	42	66	300
128	28	44	200
320	70	110	500

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
200	192	8	64	0·333
120	128	− 8	64	0·500
40	42	− 2	4	0·095
30	28	2	4	0·143
60	66	− 6	36	0·545
50	44	+ 6	36	0·818
				$\Sigma(O-E)^2/E = 2·434$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2·434$$

Degrees of freedom $= v = (c-1)(r-1) = (3-1)(2-1) = 2$

For $v = 2$, $\chi^2_{·05} = 5·99$

Since, the calculated value of χ^2 is less than the table value, we accept the hypothesis and conclude that the drug is effective.

Example 30 : Fit a Poisson Distribution and test the goodness of fit from the following data:

No. of mistakes per page	0	1	2	3	4	Total
No. of pages	211	90	19	5	0	325

(Given $e^{-0·44} = 0·6440$)

Solution. (i) Fitting of Poisson Distribution

Mistake (X)	Pages (f)	fX
0	211	0
1	90	90
2	19	38
3	5	15
4	0	0
	$\Sigma f = 325$	$\Sigma fX = 143$

$$\therefore \quad \bar{X} = m = \frac{\Sigma fx}{\Sigma f} = \frac{143}{325} = \cdot 44$$

By Poisson distibution, the expected frequency (number) of pages containing x mistakes is given by :

$$f(X) = N \,.\, P(X) = 325 \times \frac{e^{-\cdot 44} \times (\cdot 44)^x}{x!}$$

Also $P(0) = e^{-\cdot 44} = \cdot 6440$ Computation of Expected Frequencies

Computation of Expected Frequencies

X	$fe(x) = N \times P(x)$	E.
0	$f(0) = 325 \times e^{-44} = 325 \times \cdot 64470$	$= 209 \cdot 30$
1	$f(1) = f(o) \times \frac{m}{1} = 209 \cdot 30 \times \cdot 44$	$= 92 \cdot 09$
2	$f(2) = f(1) \times \frac{m}{2} = 92 \cdot 14 \times \frac{\cdot 44}{2}$	$= 20 \cdot 26$
3	$f(3) = f(2) \times \frac{m}{3} = 20 \cdot 26 \times \frac{\cdot 44}{3}$	$= 2 \cdot 97$
4	$f(4) = f(3) \times \frac{m}{4} = 2 \cdot 97 \times \frac{\cdot 44}{4}$	$= \cdot 3267$

(b) Test of Goodness of Fit : Let us teke the hypothesis that there is no difference between the observed and expected frequencies. Since, the frequency at one corner are less than 5, they would be combined with the adjacent frequency :

O	E	(O – E)	$(O-E)^2$	$(O-E)^2 / E$
211	209.3	+ 1·7	2·89	·0138
90	92.09	– 2·09	4·368	0·0474
19	20.26	– 1·26	1.587	0·078
5 0	3.29	+ 1·71	2·924	0·88
				$\Sigma(O-E)^2 / E = 1.0192$

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = 1 \cdot 0192$$

Degrees of freedom $(v) = n - 2 = 4 - 2 = 2$

$\chi^2_{.05}$ for 2 d.f. $= 5 \cdot 99$.

Since, the calculated value of χ^2 is less than the table of χ^2, we accept the null hypothesis and conclude that the fit is good.

Example 31 : Four coins are tossed 160 times and the following results were obtained :

No. of heads	0	1	2	3	4
Observed frequency	17	52	54	31	6

Under the asumption that coins are unbiased, find the expected frequencies of getting 0, 1, 2, 3, or 4 heads and test the goodness of fit.

Solution. On the assumption that the coins are undiased, the expected frequencies of getting 0, 1, 2, 3, and 4 heads will be given by the formula of binomial distribution:

$$f(X) = N.P(X) = N.\,{}^{n}c_{x} \cdot q^{n-x} \cdot p$$

Here, $p = P(H) = 1/2$, $q = P(T) = 1 - \frac{1}{2} = \frac{1}{2}$, $n = 4$, $N = 160$

No. of Heads	$fe(X) = N \times P(X)$	E.
0	$160 \times {}^{4}C_{0} \left(\frac{1}{2}\right)^{4} \cdot \left(\frac{1}{2}\right)^{0}$	= 10
1	$160 \times {}^{4}C_{1} \left(\frac{1}{2}\right)^{3} \cdot \left(\frac{1}{2}\right)^{1}$	= 40
2	$160 \times {}^{4}C_{2} \left(\frac{1}{2}\right)^{2} \cdot \left(\frac{1}{2}\right)^{2}$	= 60
3	$160 \times {}^{4}C_{3} \left(\frac{1}{2}\right)^{1} \cdot \left(\frac{1}{2}\right)^{3}$	= 40
4	$160 \times {}^{4}C_{4} \left(\frac{1}{2}\right)^{0} \cdot \left(\frac{1}{2}\right)^{4}$	= 10

(b) Test of Goodness of Fit : Let us take the hypothesis that there is no difference between the observed frequencies and expected freqeucies.

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
17	10	7	49	4·900
52	40	12	144	3·600
54	60	– 6	36	0·600
31	40	–9	81	2·025
6	10	– 4	16	1·600
				$\Sigma(O-E)^2/E = 12.725$

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 12{\cdot}725$$

Degrees of freedom $(v) = n - 1 = 5 - 1 = 4$

Tabulated value of $\chi^2_{.05}$ for 4 d.f. $= 9{\cdot}49$.

Since, the calculated value of χ^2 is greater than the table value, we reject the hypothesis and hence conclude that the fit is poor.

Example 32 : Given the following actual and theoretical frequencies, test the goodness of fit:

Actual Frequency	25	50	75	102
Theoretical Frequency	36	54	72	90

Solution. Let us take the hypothesis that there is no difference in the actual frequencies and theoretical frequencies.

Computation of χ^2

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
25	36	− 11	121	3·361
50	54	− 4	16	0·296
75	72	+ 3	9	0·125
102	90	+ 12	144	1·600
				$\Sigma(O-E)^2/E = 5.382$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 5 \cdot 382$$

Degrees of freedom $(v) = n - 1 = 4 - 1 = 3$

For $v = 3, \chi^2_{0.05} = 7 \cdot 815$.

Since, the calculated value of χ^2 is less than the table value, we accept the hypothesis and therefore conclude that the there is no difference in the actual frequencies and theoretical frequencies *i.e.* the fit is good.

Example 33. **In a certain sample of 2000 families, 1400 families are consumers of tea. Out of 1800 Hindu families, 1236 families consume tea. Use Chi-square test to test whether there is any significant difference between the consumption of tea among Hindu and Non-Hindu families. Use 5% level of significance.**

Solution. The above data can be conveniently arranged in the following table as :

	Hindu	Non-Hindu	Total
No. of families consuming tea	1236	164	1400
No. of families not consuming tea	564	36	600
Total	1800	200	2000

Let the null hypothesis be that there is no significant difference between the consumers of tea among Hindu and Non-Hindu families or that the two attributes (consmption of tea and community) are independent.

On the basis of this hypothesis, the expected frequencies are.

$$E_{11} = \frac{1800 \times 1400}{2000} = 1260$$

Te remaining frequencies are found out by subtracting from the column and row totals.

The expected frequencies table would be as follows:

1260	140	1400
540	60	600
1800	200	2000

Applying χ^2-test :

O	E	$(O-E)^2$	$(O-E)^2/E$
1236	1260	576	0·457
564	540	576	1·067
164	140	576	4·114
36	60	576	9·600
			$\Sigma(O-E)^2/E = 15·238$

$$\therefore \quad X^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 15·238$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{·05} = 3·84$

Since, the calculated value of χ^2 is much greater than the table value of χ^2, we reject the null hypothesis and conclude that the two communities differ significantly as regard to the consumption of tea among them.

IMPORTANT POINTS

Chi-square (χ^2) test is used for :

(i) Testing the Significance of Population Variance

χ^2-test is used to test the signficiance of the population variance. The significance is tested by using the formula :

$$\chi^2 = \frac{\Sigma(x-\bar{x})^2}{\sigma^2} \quad \text{or} \quad \frac{ns^2}{\sigma^2} \quad \text{or} \quad \frac{(n-1)\,s_1^2}{\sigma^2}$$

Degrees of freedom $v = n-1$

Now the calculated value of χ^2 > tabulated value of χ^2, we reject the null hypothesis H_0. Otherwise, we accept H_0.

(ii) Testing the independence of attributes in a contingency table of order $r \times c$.

In case of contingency table. we set up the hypothesis that the two attributes are independent and on the basis of this assumption, we calculate expected frequency of each cell with the following formula :

$$\text{Expected frequency (E)} = \frac{\text{Total of row in which it occurs} \times \text{Total of column in which it occurs}}{\text{Total no. of observations}}$$

and finally we calculate

$$X^2 = \frac{\Sigma(O-E)^2}{E}$$

Degrees of freedom $= (r-1)(c-1)$

Now, if calculated value of χ^2 < tabulated value of χ^2 at 5% level of significance for $(r-1)(c-1)$ d.f., we accept our hypothesis otherwise reject it and conclude accordingly.

(iii) Testing the goodness of fit.

χ^2-test is used in testing the hypothesis that the observed sample distribution agrees with the theoretical distribution *i.e.*, there is no difference between the observed and expected frequencies. The significance of the difference between observed and expected frequencies are tested as follows :

Given that :

O : O_1, O_2, O_3, ..., O_n

E : E_1, E_2, E_3, ..., E_n

We calculate :

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$$

$$\text{d.f.} = k - 1$$

Now if the calculated value of χ^2 < table of χ^2 for $(k-1)$, d.f., then we accept the hypothesis and conclude accordingly.

QUESTIONS

1. Describe the χ^2-test of significance and state the various uses to which it can be put.
2. (a) What is χ^2-test of goodness of fit ? What precautions are necessary while using this test ?
   (b) What is Chi-square test of independence ? Under what conditions is it applicable ?
3. What is χ^2-test ? Give various uses of χ^2-test. What are the limiting values of χ^2 ? How will you determine the degrees of freedom for χ^2 -test ?
4. Discuss the uses of χ^2-test.
5. Discuss the precautions which should be kept in mind while using χ^2-test.

6

F-Test and Analysis of Variance

INTRODUCTION

Analysis of Variance (abbreviated as ANOVA) is one of the most powerful techniques of statistical analysis. It was developed by **R.A. Fisher.** Initially, this technique was used in agricultural experiments but now a days it is widely used in natural, social and physical science. This technique is used to test whether the difference between the means of three or more populations is significant or not. By using the technique of analysis of variance, we can test whether the different varieties of seeds or fertilisers applied on different plots of land differ significantly or not as regard their average yields. A manager of a firm may use this technique to test whether there is significant difference in the average sale figures of different salesmen employed by the firm. Analysis of variance thus enables us to test on the basis of sample observations whether the means of three or more population is significantly different or not.

MEANING OF ANALYSIS OF VARIANCE

Analysis of variance is a statistical technique with the help of which the total variation of the data is split up into various components which may be attributed to various "sources" or "causes" of variation. There may be variation between the samples and also within the samples. By comparing the variance between the samples and variance within samples, analysis of variance helps in testing the homogenity of several population means. In the words of **Yule and Kendell, "The analysis of variance is essentially a procedure for testing the difference between different groups of data for homogenity" To quote R.A. Fisher,"Analysis of variance is the separation of the variance ascrible to one group of causes from the variance ascrible to other groups."** Thus, the analysis of variance obtains a measure of the variance within the samples and also variance between the samples and then test the significance of the difference between the means of two or more populations.

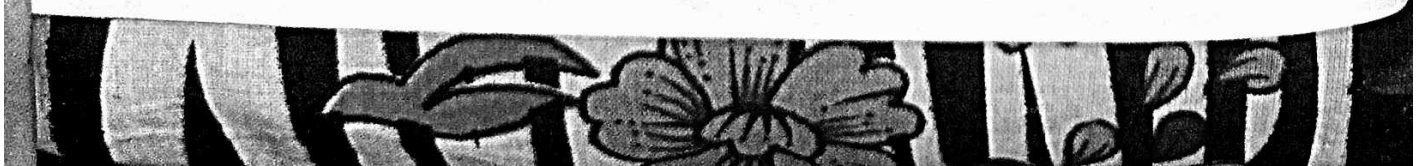
ASSUMPTIONS OF ANALYSIS OF VARIANCE

The underlying assumptions for the study of analysis of variance are :

(1) **Normal Population :** All the population from which samples have been drawn are normally distributed.

(2) **Independence of Samples :** The samples are randomly and independently drawn from the population. That is, each of the sample is independent of the other samples.

(3) **Same Population Variance :** The population from where the samples have been taken should have equal variance ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma^2$, say) where σ^2 is unknown.



(4) **Additivity :** The sum of variances of all the components should be equal to the total variance.

The analysis of variance technique is valid under the above assumptions. Otherwise the results will be illusory with less importance.

USES AND UTILITY OF ANALYSIS OF VARIANCE

The following are some of the uses of analysis of variance :

(1) **Test of the significance between the means of several samples :** The analysis of variance is used to test the hypothesis whether the means of several samples are significantly different or not.

(2) **Test of the significance between the variance of two samples :** F-ratio in the analysis of variance is used to test the significance of the difference between the variance of two samples.

(3) **Study of homogenity in case of two-way classification :** Homogenity of data can also be studied in analysis of variance of two-way classification because in this case the data are classified into different parts on two bases.

(4) **Test of correlation and regression :** The analysis of variance is used to test the significance of multiple correlation coefficient. The linearity of regression is also tested with its help.

TECHNIQUE OF ANALYSIS OF VARIANCE

The technique of analysis of variance is studied under the following two headings "

(A) One way Classification, and

(B) Two way Classification.

(A) One way Classification : In one-way classification, the data are classified on the basis of one factor or criterion only. For example, the yields of several plots of land may be classified according to different types of seeds, fertilisers, etc. In case of one-way classification, the analysis of variance can be done by the following methods :

(1) Direct Method

(2) Short-cutn Method

(3) Coding Method

(1) Direct Method : Under direct method, the following steps are followed :

(i) **Null Hypothesis ,** $H_0 : \mu_1 = \mu_2 = ... \mu_k$ *i.e.*, the means of the population from which the samples have been taken are equal and there is no difference among them.

(ii) **Variance between the samples :** Compute the mean ($\bar{x}$) of each sample. Find the combined mean ($\bar{\bar{x}}$) of all the sample means. Take deviations from $\bar{\bar{x}}$ *i.e.*, compute $\bar{x} - \bar{\bar{x}}$ and then square these deviations $(\bar{x} - \bar{\bar{x}})^2$. Find the sum of these squared deviations and divide it by the corresponding degrees of freedom $(k-1)$, where k is the number of samples. Thus, we find the variance between the samples. Symbolically,

Sum of squares of the deviations between samples (SSB)

$$= n_1 (\bar{x}_1 - \bar{\bar{x}}_2)^2 + n_2 (\bar{x}_1 - \bar{\bar{x}}_2)^2 \ ... + n_k (\bar{x}_k - \bar{\bar{x}})^2$$

Degrees of freedom, $v_1 = k - 1$

$\therefore$ Variance between samples = (MSB) $= \dfrac{SSB}{k-1}$

(iii) Variance within samples : Take the deviations in each sample from the respective sample means, $x_1 - \bar{x}_1, x_2 - \bar{x}_2$ ... and find their squares, $(x_1 - \bar{x}_1)^2, (x_2 - \bar{x}_2)^2$. Divide the sum of these squares of deviations by relevant degrees of freedom $v_2 = N - k$, where N is the total number of observations. Thus, we find the variance with samples. Symbolically,

Sum of squares of the deviation within samples (SSW)

$$= (x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2 + \ldots + (x_k - \bar{x}_k)^2$$

Degrees of freedom, $v_2 = N - k$

Variance within the samples (MSW) $= \dfrac{SSW}{N-k}$

(iv) Analysis of Variance Table : The results of the above calculations is presented in a table, called **Analysis of variance** or ANOVA table as follows :

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of squares (MSS)	F-ratio
Between Samples	$\Sigma n_k \left(\bar{x}_k - \bar{\bar{x}}\right)^2$ (SSB)	$k-1$	$\dfrac{SSB}{k-1} = MSB$	$F = \dfrac{MSB}{MSW}$
Within Samples	$\Sigma\left(x - \bar{x}_k\right)^2$ (SSW)	$N-k$	$\dfrac{SSW}{N-k} = MSW$	
Total	$\Sigma\left(x - \bar{\bar{x}}\right)^2$ (TSS)	$N-1$		

(v) The calculated value of F is compared with the table value of F for $(k-1, N-k)$ d.f. at a specified level of significance. If the calculated value of F is less than the table value of F, we accept the null hypothesis and conclude that all population means are equal, otherwise they may be taken to be unequal.

The following examples illustrate the procedure involved under direct method :

Example 1. **Three varieties *A, B* and *C* of wheat are sown in four plots each and the following yields per acre were obtained :**

Plots	Varieties		
	A	*B*	*C*
1	8	7	12
2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Set up a table of analysis of variance and find out where there is a significant difference between the mean yields of these varieties. (Given $F_{.05} = 4{\cdot}26, 3{\cdot}38$ and $3{\cdot}88$ at d.f. (9, 2) (3, 9) and (2, 12) respectively).

Solution. Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ *i.e.*, mean yields of three varieties are the same.

Computation of Arithmetic Mean

X_1	X_2	X_3
8	7	12
10	5	9
7	10	13
14	9	12
11	9	14
$\Sigma X_1 = 50\ n_1 = 5$	$\Sigma X_2 = 40\ n_2 = 5$	$\Sigma X_3 = 60\ n_3 = 5$

$$\bar{x}_1 = \frac{50}{5} = 10, \quad \bar{x}_2 = \frac{40}{5} = 8, \quad \bar{x}_3 = \frac{60}{5} = 12$$

Grand Mean, or $$\bar{\bar{x}} = \frac{10+8+12}{3} = 10$$

Variance between Samples

Sum of squares of the deviations between samples (SSB)

$$SSB = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2$$
$$= 5(10-10)^2 + 5(8-10)^2 + 5(12-10)^2$$
$$= 5\times 0 + 5\times 4 + 5\times 4 = 40$$

Degrees of freedom, $v_1 = k - 1 = 3 - 1 = 2$

Variance between samples (MSB) $= \frac{SSB}{k-1} = \frac{40}{2} = 20$

Variance within Samples

X_1	$(X_1 - \bar{x}_1)$	$(X_1 - \bar{x}_1)^2$	X_2	$(X_2 - \bar{x}_2)$	$(X_2 - \bar{x}_2)^2$	X_3	$(X_3 - \bar{x}_3)$	$(X_3 - \bar{x}_3)^2$
8	− 2	4	7	− 1	1	12	0	0
10	0	0	5	− 3	9	9	− 3	9
7	− 3	9	10	2	4	13	+ 3	1
14	+ 4	16	9	1	1	12	+ 0	0
11	+ 1	1	9	1	1	14	+ 2	4
$\bar{x}_1 = 10$		$\Sigma(X_1 - \bar{x}_1)^2 = 30$	$\bar{x}_2 = 8$		$\Sigma(X_2 - \bar{x}_2)^2 = 16$	$\bar{x}_3 = 12$		$\Sigma(X_3 - \bar{x}_3)^2 = 14$

Sum of the squares of the deviations within samples (SSW)

$$SSW = \Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2 + \Sigma(x_3 - \bar{x}_3)^2$$
$$= 30 + 16 + 14 = 60$$

Degrees of freedom, $v_2 = N - k = 15 - 3 = 12$.

Variance with samples (MSW) $= \frac{SSW}{N-k} = \frac{60}{12} = 5$

The results of the above calculation is presented in a table called ANOVA table as follows :

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of squares (MSS)	F-Ratio
Between Samples	40 (SSB)	2	$\frac{40}{2} = 20$ (MSB)	$F = \frac{20}{5} = 4$
Within Samples	60 (SSW)	12	$\frac{60}{12} = 5$ (MSW)	
Total	100 (TSS)	14		

For $v_1 = 2, v_2 = 12$, the table value of F at 5% level os significance is 3·88. Since, the computed value of F is greater than the table value *i.e.*, $F > F_{0.05}$, we reject null hypothesis and conclude that the difference between the mean yields of 3 varieties is significant.

(2) Short cut Method : The direct method is much calculative and time consuming and moreover, the calculation becomes more complicated when the arithmetic mean is not in whole number. In such a case, short-cut method is used. It involves the following steps :

(i) Find the sum of all sample observations and their squares :

Sum of the sample values $= \Sigma X_1, \ \Sigma X_2, \Sigma X_3, \ldots \Sigma X_k$

Sum of the squares of sample values $= \Sigma X_1^2, \Sigma X_2^2, \Sigma X_3^2, \ldots \Sigma X_k^2$

(ii) Find the correction factor : To obtain the correction factor, divide the square of the total of all values by the number of values *i.e.*,

$$C.F. = \frac{T^2}{N}$$

where, $C.F.$ = Correction factor, T^2 = Square of the total units of the samples

N = Total no. of units of the samples

(iii) Find the total sum of squares, TSS : To find the total sum of squares subtract correction factor from the sum of the squares of all samples values *i.e.*,

$$TSS = (\Sigma X_1^2 + \Sigma X_2^2 + \ldots \Sigma X_k^2) - \frac{T^2}{N}$$

(iv) Find the sum of squares between samples, SSB : To find the SSB, divide the sum of the squares of each samples by their size and then find their sum. Subtract the correction factor from this sum *i.e.*,

$$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \ldots + \frac{(\Sigma X_k)^2}{n_k}\right] - \frac{T^2}{N}$$

(v) Find the sum of squares within samples, SSW : It is obtained by deducting the sum of squares between the samples from the total sum of squares *i.e.*,

$$SSW = TSS - SSB$$

(vi) Analysis of Variance Table and Interpretation of Significance : Analysis of variance table and interpretation are the same as in case of direct method.

Note : In case sample sizes are unequal, there is no change in the analysis of variance. Utmost care must be taken while calculating degrees of freedom in such cases.

xample 2. **Three varieties *A*, *B* and *C* of wheat are sown in four plots each and the following yields per acre were obtained.**

Plots	Varieties		
	A	*B*	*C*
1	8	7	12
2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Is there any significant difference in the production of three varieties ? Use short cut method.

lution. Null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$ *i.e.*, there is no difference between the mean yield of three varieties.

A		*B*		*C*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
8	64	7	49	12	144
10	100	5	25	9	81
7	49	10	100	13	169
14	196	9	81	12	144
11	121	9	81	14	196
$\Sigma X_1 = 50$ $n_1 = 5$	$\Sigma X_1^2 = 530$	$\Sigma X_2 = 40$ $n_2 = 5$	$\Sigma X_2^2 = 336$	$\Sigma X_3 = 60$ $n_3 = 5$	$\Sigma X_3^2 = 734$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 50 + 40 + 60 = 150$$

Correction Factor, $$C.F. = \frac{T^2}{N} = \frac{(150)^2}{15} = 1500$$

TSS = Total sum of squares

$$= (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2) - C.F.$$

$$= 530 + 336 + 734 - 1500 = 100$$

SSB = Sum of squares between samples

$$= \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$

$$= \left[\frac{(50)^2}{5} + \frac{(40)^2}{5} + \frac{(50)^2}{5}\right] - 1500$$

$$= \frac{1}{5} \cdot [2500 + 1600 + 3600] - 1500 = 1540 - 1500 = 40$$

SSW = TSS – SSB = 100 – 40 = 60.

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of square (MSS)	F-Ratio
Between samples	40 (SSB)	2	20	$F = \frac{20}{5} = 4$
Within Samples	60 (SSW)	12	5	
Total	100 (TSS)	14		

For $v_1 = 2, v_2 = 12$, the table value of F at 5% level of significance is 3·88. Since the calculated value of F is greater than the table value *i.e.*, $F > F_{0.05}$, we reject the null hypothesis and hence conclude that the difference between the mean yields of three varieties is significant.

(3) Coding Method : The short-cut method becomes tedious when the magnitude of the given values is large. Coding method simplified the calculations involved in the short-cut method and is popularly used in practice. Coding refers to the addition, subtraction, multiplication or division of data by a constant quantity. As the F-statistic in the analysis of variance is a ratio, its value does not change if all the given values are coded *i.e.*, either multiplied or divided by a common factor or if a common figure is either subtracted or added to each of the given values. By this method, big figures are reduced in magnitude by subtraction or division and the work in simplified without altering the value of F. Analysis of variance table and interpretation are the same as in case of short-cut method.

The following examples illustrate the coding method :

Example 3 **The following table gives the yields of four varieties of wheat grown in 2 plots:**

Plots	Varieties			
	A	*B*	*C*	*D*
1	200	230	250	300
2	190	270	300	270
3	240	150	145	180

Is there any significant difference in the production of these varieties ?

Solution. Null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ *i.e.*, there is no significant difference in the mean yield of four varieties.

In order to simplify the calculation, subtract 200 from each sample value and dividing the difference by 10.

Coded Data

A		*B*		*C*		*D*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
0	0	3	9	5	25	10	100
−1	1	7	49	10	100	7	49
4	16	−5	25	−5·5	30·25	−2	4
$\Sigma X_1 = 3$ $n_1 = 3$	$\Sigma X_1^2 = 17$	$\Sigma X_2 = 5$ $n_2 = 3$	$\Sigma X_2^2 = 83$	$\Sigma X_3 = 9·5$ $n_3 = 3$	$\Sigma X_3^2 = 155·25$	$\Sigma X_4 = 15$ $n_4 = 3$	$\Sigma X_4^2 = 153$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 3 + 5 + 9{\cdot}5 + 15 = 32{\cdot}5$$

Correction Factor, $$C.F. = \frac{T^2}{N} = \frac{(32{\cdot}5)^2}{12} = 88{\cdot}02$$

TSS = Total sum of squares

$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2] - C.F.$$

$$= [17 + 83 + 155{\cdot}25 + 153] - 88{\cdot}02 = 320{\cdot}23$$

SSB = Sum of squares between the samples

$$= \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4}\right] - C.F.$$

$$= \left[\frac{(3)^2}{3} + \frac{(5)^2}{3} + \frac{(9{\cdot}5)^2}{3} + \frac{(15)^2}{3}\right] - 88{\cdot}02 = 28{\cdot}39$$

SSW = SST − SSB = 320·23 − 28·39 = 291·84

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of square (MSS)	F-Ratio
Between samples	28·39	3	9·46	$F = \frac{36{\cdot}48}{9{\cdot}46} = 3{\cdot}85$
Within samples	291·84	8	36·48	
Total	320·23	11		

For $v_1 = 8, v_2 = 3$, the table value of F at 5% level of significance is 4·07. Since the calculated value of F is less than the table value *i.e.*, $F < F_{0{\cdot}05}$, we accept the null hypothesis and conclude that there is no significant difference in the mean yield of four varieties.

Example 4. **The following figures relate to producting in kg of three varieties *A*, *B* and *C* of wheat sown in 12 plots :**

A	14	16	18		
B	14	13	15	22	
C	18	16	19	19	20

Is there any significant difference in the production of these varieties ?

Solution. Null Hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$ *i.e.*, there is no difference in the production of three varieties.

In order to simplify the calculations, the given data are coded by subtracting 12 from each figure. The deviations and their squares are as follows :

Coded Data

A		*B*		*C*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
2	4	2	4	6	36
4	16	1	1	4	16

6	36	3	9	7	49
—	—	10	100	7	49
—	—	—	—	8	64
$\Sigma X_1 = 12$, $n_1 = 3$	$\Sigma X_1^2 = 56$	$\Sigma X_2 = 16$, $n_2 = 4$	$\Sigma X_2^2 = 144$	$\Sigma X_3 = 32$, $n_3 = 5$	$\Sigma X_3^2 = 214$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 12 + 16 + 32 = 60$$

$$C.F. = \frac{T^2}{N} = \frac{(60)^2}{12} = 300$$

TSS = Total sum of squares

$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= 56 + 144 + 214 - 300 = 84$$

SSB = Sum of squares between samples

$$= \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$

$$= \left[\frac{(12)^2}{3} + \frac{(16)^2}{4} + \frac{(32)^2}{5}\right] - 300$$

$$= 48 + 64 + 204{\cdot}8 - 300 = 16{\cdot}8$$

SSW = TSS − SSB

$$= 84 - 16{\cdot}8 = 67{\cdot}20$$

ANOVA Table

Source of Variation	Sume of squares	Degrees of freedom (d.f.)	Mean sum (MSS)	F-Statistic
Between Samples	16·8	2	8·4	$F = \frac{8{\cdot}4}{7{\cdot}467} = 1{\cdot}125$
Within samples	67·20	9	7·467	
Total	84	11	—	

For $v_1 = 2, v_2 = 9$, the table value of F at 5% level of significance is 4·261. Since, the calculated value of F is less than the table value of F. We accept the null hypothesis and conclude that there is no difference in the mean productivity of three varieties.

EXERCISE – 1

1. The following table gives the yields on 15 sample plots under three varieties of seed :

Seeds / Plots	P_1	P_2	P_3	P_4	P_5
S_1	20	21	23	16	20
S_2	18	20	17	15	25
S_3	25	28	22	28	32

Is the difference between varieties significant.
For (2, 12) degrees of freedom, $F_{0{\cdot}05} = 3{\cdot}88$.

[Ans. $F = 8{\cdot}14$, Reject H_0]

2. The following data gives the retail prices of a commodity in some shops selected at random in four cities :

City	Prices (Rs./Kg)			
A	22	24	27	23
B	20	19	23	19
C	10	17	21	18
D	24	25	29	26

Carry out the analysis of variance to test the significance of the difference between the prices of the commodities in four cities

For (3, 12) degrees of freedom, $F_{0.05} = 3.49$

For (3, 9) degrees of freedom, $F_{0.05} = 3.86$

[Ans. $F = 2.114$, Reject H_0]

3. Three varieties A, B and C of wheat were sown in four plots each and the following yields per acre were obtained :

Plots of Land	Varieties of Wheat		
	A	B	C
1	10	9	4
2	6	7	7
3	7	7	7
4	9	5	6

Set up a table of analysis of variance and find out whether there is a significant difference between the mean yield of three varieties (Given $F_{0.05} = 4.25$, 3.86 and 4.10 at d.f. (2, 9), (3, 9) and (2, 10) respectively).

[Ans. $F = 1.771 < F_{0.05}$, Accept H_0]

4. A test was given to 5 students chosen at random from M. Com. class of each of the three universities in Haryana. Their scores were found as follows :

University	Scores				
A	90	70	60	50	80
B	70	40	50	40	50
C	60	50	60	70	60

Perform analysis of variance and show if there is any significant difference between the scores of students in the three universities. [Given F value at 5% = 3.89]

[Ans. $F = 3.33$, Accept H_0]

5. The following figures relate to the number of units sold in five different areas by four salesmen :

Area	Number of units			
	A	B	C	D
1	80	100	95	70
2	82	110	90	75
3	88	105	100	82
4	85	115	105	88
5	75	90	80	65

Is there a significant difference in the efficiency of these salesmen ?

Hint : See Example 11.

Table values of $F_{0.05}$ for $v_1 = 3, v_2 = 16$ is 3·24.

[**Ans.** $F = 10{\cdot}61 > F_{0.05}$, Reject H_0 *i.e.*, there is a significant difference in the efficiency of the four salesmen]

6. The Amrit Vanaspati Company of Rajpura (Punjab) wishes to test whether its three salesmen *A*, *B* and *C* tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week of October, 2004, There have been 14 sales calls – *A* made 5 calls, *B* made 4 calls and *C* made 5 calls.

Following are the weekly sales record of the three salesmen :

A (Rs.)	300	400	300	500	000
B (Rs.)	600	300	300	400	—
C (Rs.)	700	300	400	600	500

Perform the analysis of variance and draw your conclusions.

[Given $F_{0.05}$ (2, 11) = 3·98; $F_{0.05}$ (2, 13) = 3·82] [**Ans.** $F = 1{\cdot}83$, Accept H_0]

7. Yields of 3 varieties of wheat in 3 blocks are given below :

Blocks/Varieties	1	2	3
I	200	230	300
II	190	270	270
III	240	150	180

Is the difference between varieties significant ? [**Ans.** $F = 3{\cdot}84$, Accept H_0]

8. The following figures relate to the production in Kg of three varieties *A*, *B* and *C* on wheat sown in 12 plots :

A	122	128	124	126	
B	114	116	118	114	106
C	130	128	124		

(Use Coding Method)

Is there any significant difference in the production of three varieties ?

[**Ans.** $F = 16{\cdot}90 > F_{0.05}$, Reject H_0]

9. Apply *F*-test on the following data :

X_1	X_2	X_3
25	31	24
30	39	30
36	38	28
38	42	25
31	35	28
160	185	135

(Hints : See Example 15)

Given ($F_{2, 12}$, 5 percent = 3·89) [**Ans.** $F = 7{\cdot}49$, Reject H_0]

10. To test the significance of the variation of the retail prices of a commodity in three principal cities : Bombay, Kolkata and Delhi, four shops were chosen at random in each city and prices observed in Rs. were as follows :

Bombay	16	8	12	14
Kolkata	14	10	10	6
Delhi	4	10	8	8

Do the data indicate that the prices in the three cities are significantly different ?

[Ans. $F = 2 \cdot 616$ Accept H_0]

11. To assess the significance of possible variation are in a performance in a certain test as between the grammar schools of a city, a common test was given to students take at random from the senior fifth form of each of the four schools. Carry out analysis of variance of the data and comment upon the results.

School	Marks Obtained by the students							
A	8	7	4	5	5	5	6	6
B	7	5	5	4	3	4	6	4
C	5	3	4	4	3	5	4	4
D	10	5	6	4	8	7	8	4

Given : $F_{0 \cdot 05} = 2 \cdot 95$ [Ans. $F = 5 \cdot 10$, Reject H_0]

(B) ANALYSIS OF VARIANCE IN TWO-WAY CLASSIFICATION

In two-way classification, the data are classified according to two factors. For example, the production of a manufacturing concern can be studied on the basis of workers as well as machines. A company can analyse its sales according to salesmen and seasons. In two-way classification, the following procedure is adopted in the analysis of variance :

(i) Coding method can be used to simplify the calculation.

(ii) Find the correction factor by using the formula :

$$\text{Correction Factor } (C.F.) = \frac{T^2}{N}$$

Where, T = Grand total of all the values in all the samples, N = Total number of items.

(iii) Find total sum of squares (TSS) : It is obtained by subtracting the correction factor from the total of squared values of the sample *i.e.*,

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 + \ldots \Sigma X_k^2] - \frac{T^2}{N}$$

(iv) Find the sum of squares between columns (SSC) : The total of each column is squared and divided by the number of items in the column. The correction factor is subtracted from it and SSC is obtained *i.e.*,

$$SSC = \Sigma \left(\frac{\Sigma X_c}{n_c} \right)^2 - \frac{T^2}{N}$$

Where, ΣX_c^2 = total of squared values in each column; n_c = number of items in each column.

(v) Find the sum of squares between rows (SSR) : The total of each row is squared and divided by the number of items in respective rows. The correction factor is subtracted from the total of, thus, arrived row and SSR is obtained, *i.e.*,

$$SSR = \Sigma\left(\frac{\Sigma X_r^2}{n_r}\right) - \frac{T^2}{N}$$

Where, ΣX_r^2 = Total of squared values in each row; n_r = number of items in each row.

(vi) Find the sum of the squares of the residual (SSE) : The sum of the squares of the residual is obtained by deducting the SSC and SSR from TSS. Thus

$$SSE = TSS - SSC - SSR$$

(vii) Find the number of degrees of freedom by using the formula :

No. of degrees of freedom between columns $= (c-1)$

No. of degrees of freedom between rows $= (r-1)$

No. of degrees of freedom for residual $= (c-1)(r-1)$

Total no. of degrees of freedom $= N-1$ or $cr-1$

(viii) ANOVA Table : In a two-way classification, the analysis of variance (ANOVA) table is prepared in the following way :

ANOVA Table (Two-way Classification)

Source of variation	sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between columns	SSC	$(c-1)$	$SSC \div (c-1) = MSC$	$F = \frac{MSC}{MSE}$
Between Rows	SSR	$(r-1)$	$SSR \div (r-1) = MSR$	$F = \frac{MSR}{MSE}$
Residual	SSE	$(c-1)(r-1)$	$SSE \div (c-1) \times (r-1) = MSE$	
Total	TSS	$(N-1)$ or $(cr-1)$		

(ix) Interpretation : The calculated value of F is compared with the table value of F and if the calculated value of F is greater than the table value at a specified level of significance, the null hypothesis is rejected and concluded that the difference is significant otherwise vice versa.

Example 5. **The following data represent the number of units of a commodity produced by 3 different workers using 3 different machines :**

Machine / Workers	A	B	C
X	16	64	40
Y	56	72	56
Z	12	56	28

Test (i) Whether the mean productivity is the same for the different machine types (ii) whether the three workers differ with regard to mean productivity.

Solution. Let us take the hypothesis that :

(i) The mean productivity for three different machines is the same.

(ii) Three workers do not differ with respect to their mean productivity.

Machine Workers	Data			Row Total			
	X_1	X_2	X_3		X_1^2		
X	16	64	40	120	256	409	
Y	56	72	56	184	3136	5184	
Z	12	56	28	96	144	3136	
Column Total	84	192	1245	$T = 400$	3536	12416	552

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 84 + 192 + 14 = 400$$

$$C.F. = \frac{T^2}{N} = \frac{(400)^2}{9} = 17777 \cdot 78$$

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= [3536 + 12416 + 5520] - 17777 \cdot 78$$

$$= 21472 - 17777 \cdot 78 = 3694 \cdot 22$$

$$SSC = \left[\frac{(84)^2}{3} + \frac{(192)^2}{3} + \frac{(124)^2}{3}\right] - 17777 \cdot 78\,(C.F.)$$

$$= \frac{1}{3}[7056 + 36864 + 15376] - 17777 \cdot 78$$

$$= 19765 \cdot 33 - 17777 \cdot 78 = 1987 \cdot 55$$

$$SSR = \left[\frac{(120)^2}{3} + \frac{(184)^2}{3} + \frac{(96)^2}{3}\right] - 17777 \cdot 78\,(C.F.)$$

$$= \frac{1}{3}[14400 + 33856 + 9216] - 17777 \cdot 78$$

$$= 19157 \cdot 33 - 17777 \cdot 78 = 1379 \cdot 55$$

$$SSE = TSS - SSC - SSR$$

$$= 3694 \cdot 22 - 1987 \cdot 55 - 1379 \cdot 55 = 327 \cdot 12$$

Degrees of freedom are

$$TSS = N - 1 = 9 - 1 = 8$$

$$SSC = c - 1 = 3 - 1 = 2$$

$$SSR = r - 1 = 3 - 1 = 2$$

$$SSE = (c-1)(r-1) = 2 \times 2 = 4$$

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between columns (Machines)	1987·55	2	$\frac{1987 \cdot 55}{2} = 993 \cdot 77$	$F = \frac{993 \cdot 73}{81 \cdot 78} = 12 \cdot 15$
Between Rows (Workers)	1379·55	2	$\frac{1379 \cdot 55}{2} = 689 \cdot 77$	$F = \frac{689 \cdot 77}{81 \cdot 78} = 8 \cdot 43$
Residual/Error	327·12	4	$\frac{327 \cdot 12}{4} = 81 \cdot 78$	
Total	3694·22	8		

Interpretation

For Machines

(i) For $v_1 = 2, v_2 = 4$, the table value of $F_{0.05} = 6 \cdot 94$.

Since the calculated value of F is greater than the critical value of F, the null hypothesis is rejected. Hence the mean productivity is not the same for three different machines.

For Workers

(ii) For $v_1 = 2$ and $v_2 = 4$, the table value of $F_{0.05} = 6 \cdot 94$.

Since the calculated value of F is greater than the critical value of F, the null hypothesis is rejected. Hence the workers differ with regard to mean productivity.

Example 6. **The following table gives the number of refrigerators sold by 4 salesmen in three seasons-summer, winter and monsoon :**

Season	Salesmen			
	A	*B*	*C*	*D*
Summer	62	62	32	60
Winter	46	48	52	54
Monsoon	42	46	48	48

Is there a significant difference in the sales made by the four salesmen ?
Is there a significant difference in the sale made during different seasons ?

Solution. (i) Let us take the null hypothesis that the mean sales made by the four salesmen is the same.

(ii) The sales do not differ with regard to seasons.

In order to simplify calculations, the given data is coded by subtracting 50 from each observation. The data in the coded form are given below :

Season	Coded Data				Row Total	Squares of Coded Data			
	X_1	X_2	X_3	X_4		X_1^2	X_2^2	X_3^2	X_4^2
S	+ 12	+ 12	− 18	+ 10	+ 16	144	144	324	100
W	− 4	− 2	+ 2	+ 4	+ 0	16	4	4	16
M	− 8	− 4	− 2	− 2	− 16	64	16	4	4
Column Total	0	+ 6	− 18	+ 12	$T = 0$	224	164	332	120

$$T = 0 + 6 - 18 + 12 = 0$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(0)^2}{12} = 0$$

$$\text{TSS} = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \text{C.F.}$$

$$= 224 + 164 + 332 + 120 - 0 = 840$$

$$\text{SSC} = \left[\frac{(0)^2}{3} + \frac{(6)^2}{3} + \frac{(-18)^2}{3} + \frac{(12)^2}{3}\right] - 0 \quad \text{(C.F.)}$$

$$= 0 + 12 + 108 + 48 - 0 = 168$$

$$SSR = \left[\frac{(16)^2}{4} + \frac{(0)^2}{4} + \frac{(-16)^2}{4}\right] - 0 \quad (C.F.)$$

$$= 64 + 0 + 64 - 0 = 128$$

$$SSE = TSS - SSC - SSR = 840 - 168 - 128 = 544$$

Degrees of freedom are :

$$TSS = N - 1 = 12 - 1 = 11$$

$$SSC = (c - 1) = (4 - 1) = 3$$

$$SSR = (r - 1) = (3 - 1) = 2$$

$$SSE = (c - 1)(r - 1) = 3 \times 2 = 6$$

The ANOVA table is shown as :

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean Sum of squares (MSS)	*F*-Ratio
Between Columns (Salesmen)	168	3	$MSC = \frac{168}{3} = 56$	$F = \frac{MSC}{MSE} = \frac{90 \cdot 67}{56} = 1 \cdot 619$
Between Rows (Seasons)	128	2	$MSR = \frac{128}{2} = 64$	$F = \frac{MSR}{MSE} = \frac{90 \cdot 67}{64} = 1 \cdot 4167$
Residual/Error	544	6	$MSE = \frac{544}{6} = 90 \cdot 67$	
Total	840	11		

Interpretation

(i) **For Salesmen :** The calculated value of $F = 1 \cdot 619$

Table value of F for (6, 3) d.f. $= 8 \cdot 94$

Since the calculated value of F is less than the table value of F at 5% level of significance, the null hypothesis is accepted and it can be concluded that there is no difference in the sales of the four salesmen.

(ii) **For seasons :** The calculated value of $F = 1 \cdot 4167$

Table value for (2, 6) d.f. $F_{0 \cdot 05} = 5 \cdot 14$

Since the calculated value of F is less than the table value of F at 5% level of significance, the null hypothesis is accepted and it can be concluded that all seasons are similar so far as sales is concerned.

Example 7. **Four observers determine the moisture content of samples of a powder, each man taking a sample from each of six consignments. Their assessments are given below :**

Observers	Consignments					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Analysis the data and discuss whether there is any significant difference between consignments or between observers.

(Given that $F_{0.05}^{(3, 15)} = 3.29$, $F_{0.05}^{(5, 15)} = 2.90$)

Solution. Let us take the hypothesis that :

(a) There is no significant difference between consignments.

(b) There is no significant difference between observers.

In order to simplify the calculations, the given data are coded by subtracting 10 from each figure. The deviations and their squares are as follows :

Observers	Coded Data						Row Total	Squares of Coded Data					
	X_1	X_2	X_3	X_4	X_5	X_6		X_1^2	X_2^2	X_3^2	X_4^2	X_5^2	X_6^2
1	−1	0	−1	0	1	1	0	1	0	1	0	1	1
2	2	1	−1	1	0	0	3	4	1	1	1	0	0
3	1	0	0	2	1	0	4	1	0	0	4	1	0
4	2	3	1	3	2	0	12	4	9	1	16	4	0
Column Total	4	4	−1	7	4	1	$T = 19$	10	10	3	21	6	1

$$T = 4 + 4 - 1 + 7 + 4 + 1 = 19$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(19)^2}{24} = 15.04$$

$$\text{TSS} = [10 + 10 + 3 + 21 + 6 + 1] - 15.04 \quad (\text{C.F.}) = 35.96$$

$$\text{SSC} = \left[\frac{(4)^2}{4} + \frac{(4)^2}{4} + \frac{(-1)^2}{4} + \frac{(7)^2}{4} + \frac{(4)^2}{4} + \frac{(1)^2}{4}\right] - 15.04\,(\text{C.F.}) = 9.71$$

$$\text{SSR} = \left[\frac{(0)^2}{6} + \frac{(3)^2}{6} + \frac{(4)^2}{6} + \frac{(12)^2}{6}\right] - 15.04\,(\text{C.F.}) = 13.13$$

$$\text{SSE} = \text{TSS} - \text{SSC} - \text{SSR} = 35.96 - 9.71 - 13.13 = 13.12$$

The ANOVA table is shown as :

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between columns (Between consignments)	9.71	$6 - 1 = 5$	$\frac{9.71}{5} = 1.94$	$F = \frac{1.94}{0.87} = 2.23$
Between Rows (Between Observers)	13.13	$4 - 1 = 3$	$\frac{13.13}{3} = 4.38$	$F = \frac{4.38}{0.87} = 5.03$
Residual/Error	13.12	$23 - 8 = 15$	$\frac{13.12}{15} = 0.87$	
Total	35.96	$24 - 1 = 23$		

Interpretation

(i) Between Consignments : The calculated value of $F = 2 \cdot 23$

Table value of F of (5, 15) d.f. at 5% l.o.s. $= 2 \cdot 90$

Since, the calculated value of F is less than the table value of F at 5% l.o.s., the null hypothesis is accepted and it can be concluded that there is no difference between consignments.

(ii) Between Observers : The calculated value of $F = 5 \cdot 03$

Table value of F for (3, 15) d.f. at 5% l.o.s. $= 3 \cdot 29$

Since, the calculated value of F is greater than the table value of F at 5% l.o.s., the null hypothesis is rejected and it can be concluded that there is significant difference between the observers.

Example 8. **Perform a two way ANOVA on the data given below :**

Plots of Land	Treatment			
	A	B	C	D
P	45	40	38	37
O	43	41	45	38
R	39	39	41	41

(Use coding method subtracting 40 from the given numbers)

Solution. Let us take the null hypothesis that there is no significant difference in the treatment and plots of land. By subtracting 40 from the given numbers, the deviations and their squares are given below :

Plots	Coded Data				Row Total	Squares of Coded Data			
	X_1	X_2	X_3	X_4		X_1^2	X_2^2	X_3^2	X_4^2
P	5	0	− 2	− 3	0	25	0	4	9
Q	3	1	5	− 2	7	9	1	25	4
R	− 1	− 1	1	1	0	1	1	1	1
Column Total	7	0	4	− 4	$T = 7$	35	2	30	14

$$T = 7 + 0 + 4 - 4 = 0$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{7^2}{12} = \frac{49}{12} = 4 \cdot 083$$

$$\text{TSS} = 35 + 2 + 30 + 14 - 4 \cdot 083 \text{ (C.F.)}$$

$$= 81 - 4 \cdot 083 = 76 \cdot 917$$

$$\text{SSC} = \left[\frac{(7)^2}{3} + \frac{(0)^2}{3} + \frac{(4)^2}{3} + \frac{(-4)^2}{3}\right] - 4 \cdot 083 \text{ (C.F.)}$$

$$= 22 \cdot 917 .$$

$$SSR = \frac{(0)^2}{4} + \frac{(7)^2}{4} + \frac{(0)^2}{4} - 4 \cdot 083 = 8 \cdot 167$$

$$SSE = TSS - SSC - SSR = 76 \cdot 917 - 22.917 - 8 \cdot 167 = 45 \cdot 333$$

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mans Sum of squares (MSS)	F-Ratio
Between Columns (Between Treatment)	22·917	3	$\frac{22 \cdot 913}{3} = 7 \cdot 639$	$F = \frac{7 \cdot 639}{7 \cdot 639} = 1$
Between Rows (Between Fields)	8·167	2	$\frac{8 \cdot 167}{2} = 4 \cdot 083$	$F = \frac{7 \cdot 639}{4 \cdot 083} = 1 \cdot 87$
Residual/Error	45·833	6	$\frac{45 \cdot 833}{6} = 7 \cdot 639$	
Total	76·917	11		

Interpretation

(i) Between Treatment : The calculated value of $F = 1$

Table value of F for (3, 6) d.f. at 5% l.o.s. $= 4 \cdot 76$

Since, the calculated value of F is less than the table value of F, the null hypothesis is accepted. Hence, there is no significant difference between the treatments.

(ii) Between Plots of land : The calculated value of $F = 1 \cdot 87$

Table value of F for (6, 2) d.f. at 5% l.o.s. $= 19 \cdot 3$

Since, the calculated value of F is less than the table value of F, the null hypothesis is accepted. Hence, there is no significant difference between plots of land.

EXERCISE – 2

1. A company appoints 4 salesmen A_1, A_2, A_3 and A_4 and observes their sales in three seasons : Summer, Winter and Monsoon. The figures (in lakhs) are given ahead :

Salesmen

Seasons	A_1	A_2	A_3	A_4	Total
Summer	5	4	4	7	20
Winter	7	8	5	4	24
Monsoon	9	6	6	7	28
Salesmen Total	**21**	**18**	**15**	**18**	**72**

Carry out Two-way Analysis of Variance.

[**Ans :** F Between Salesmen = 1·335, F Between Seasons = 1·498, In both the cases, H_0 is accepted]

2. Perform a two-way ANOVA on the data given below :

Plots of Land	Treatments			
	A	B	C	D
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

(Using coding method subtracting 40 from given numbers)
Given for (3, 6) d.f. F05 = 4·76
and for (2, 6) d.f. F·05 = 5·14 . [**Ans :** F Between the column = 1·312, H_0 is accepted; F Between the Rows = 1·218, H_0, is accepted]

3. The price of a certain commodity was ascertained in each of the four towns A, B, C and D in four quarters of a year. The prices are given below. Are the variations in prices between different towns and in different season significant ?

Quarters	Towns			
	A	B	C	D
I	60	50	60	50
II	50	40	65	50
III	45	35	45	50
IV	65	45	60	70

[**Ans :** F Between the column = 4·89, F between Season's = 5·00 In both the cases, H_0 is rejected]

4. The following table gives the number of units of production per day turned out by four different types of machines.

Employee	Type of Machines			
	M_1	M_2	M_3	M_4
E_1	40	36	45	30
E_2	38	42	50	41
E_3	36	30	48	38
E_4	46	47	52	44

Using analysis of variance (i) test the hypothesis that the mean production is the same for the four machines, and (ii) test the hypothesis that the employees do not differ with respect to mean production. [**Ans :** F Between Machines= 9·27, F Between Employees = 8·27, In both the cases, H_0 is rejected]

5. The following data represent the number of a commodity produced by 3 different workers using 3 different machines :

Workers	Machines		
	A	B	C
X	8	32	20
Y	28	36	38
Z	6	28	14

Test (i) whether the mean productivity is the same for different machines types, (ii) whether the three workers differ with respect to mean productivity.
[Ans : F Between Machines = 9·38, H_0 is rejected, F Between Workers = 10·31, in both cases, H_0 is rejected]

6. To study the performance of three detergents and three different water temperatures, the following whiteness readings were obtained with specially designed equipment :

Water temp.	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two way analysis of variance using 5% level of significance (Given $F_{.05} = 6·94$)
[Hint : See Example 14]
[Ans : F Between Column = 9·845, H_0 is rejected. F Between Rows = 2·381, in both the cases, H_0 is accepted]

7. You are given the following data :

Workers	Machine Type			
	A	B	C	D
1	44	35	48	38
2	48	40	50	44
3	37	38	40	36
4	45	34	45	32
5	40	44	50	40

Discuss whether there is a significant difference in mean productivity between machine types or workers. [Ans : F (Between columns) = 7·85; F (between rows) = 3·74, in both cases H_0 is rejected]

8. The following table gives the number of refrigerators sold by 4 salesmen in three months, May, June and July :

Month	Salesmen			
	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

Is there a significant difference in the sales made by the four salesmen ?
Is there a significant difference in the sales made during different months ?
[Ans : F Between columns = 1·018; F Between Rows = 3·33. In both cases, H_0 is accepted]

9. The following data represent the sales (Rs. 1000) per month of three brands of a detergent related among three cities :

Cities	Detergent A	Detergent B	Detergent C
	A	B	C
I	12	48	30
II	42	54	57
III	9	42	21

Test whether the (i) mean sales of the three brands are equal and (ii) the mean sales of detergent in each city are equal.

[Ans : F Between brands = 9·4, F between Cities = 10·3. In both cases, H_0 is rejected]

MISCELLANCEOUS SOLVED EXAMPLES

Example 9 : **To test the significance of the variations of the retail prices of a commodity in three principle cities : Bombay, Kolkata and Delhi, four shops were chosen at random in each city and prices observed in rupees were as follow :**

Bombay	16	8	12	14
Kolkata	14	10	10	6
Delhi	4	10	8	8

Do the data indicate the prices in the three cities are significantly different?

Solution : $H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.*, the mean prices in the three cities are the same.

In order to simipify the calculation, subtract 10 from each observation. The deviations and their squares are as follow :

Coded Data

Bombay		Kolkata		Delhi	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
6	36	4	16	– 6	36
– 2	4	0	0	0	10
2	4	0	0	– 2	4
4	16	– 4	16	– 2	4
$\Sigma X_1 = 10$	$\Sigma X_1^2 = 60$	$\Sigma X_2 = 0$	$\Sigma X_2^2 = 32$	$\Sigma X_3 = -10$	$\Sigma X_3^2 = 44$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10 + 0 - 10 = 0$$

$$C.F. = \frac{T^2}{N} = \frac{(0^2)}{12} = 0$$

TSS = Total sum of squares

$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= [60 + 32 + 44] - 0$$

$$= 136$$

$$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$

$$= \left[\frac{(10)^2}{4} + \frac{(0)^2}{4} + \frac{(-10)^2}{4}\right] - 0 = 50$$

SSW = SST - SSB

= 136 - 50 = 86

The various sum of squares (S.S.) along with the degrees of freedom (d.f.) are shown in the following table

ANOVA Table

Source of variation	Sum of square (S.S.)	Degrees of freedom	Mean Sum of squares (MSS)	F-Radio
Between City	50	3 – 1 = 2	25	$F = \frac{25}{9.556} = 2.616$
Within city	86	9	9·556	
Total	136	12 – 1 = 11		

For $v_1 = 2$ and $v_2 = 9$, the table value of F at 5% I. o.s. = 4·261

Since the calculated value of F is less than the table value of F the null hypothesis is accepted. We thus conclude that the mean prices in the three cities is not significantly different.

Example 10 : **Three samples, each of size 5, were chosen from three uncorrelated normal population with equal variances. Test the hypothesis that the population means are equal at 5% level.**

Sample 1	Sample 2	Sample 3
10	9	14
12	7	11
9	12	15
16	11	14
13	11	16

Solution : Let us take the hypothesis that the population means are equal for three samples *i.e.* $H_0 : \mu_1 = \mu_2 = \mu_3$

In order to simplify the calcuation, surtract 10 from each observation. The deviations and their squares are as follows :

Coded Data

Sample 1		Sample 2		Sample 3	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
0	0	– 1	1	4	16
2	4	– 3	9	1	1
– 1	1	2	4	5	25
6	36	1	1	4	16
3	9	1	1	6	36
$\Sigma X_1 = 10$	$\Sigma X_1^2 = 50$	$\Sigma X_2 = 0$	$\Sigma X_2^2 = 16$	$\Sigma X_3 = 20$	$\Sigma X_3^2 = 94$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10 + 0 + 20 = 30$$

$$C.F. = \frac{T^2}{N} = \frac{(30^2)}{15} = 60$$

TSS = Total sum of squares

$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= [50 + 16 + 94] - 60$$

$$= 160 - 60 = 100$$

$$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$

$$= \left[\frac{(10)^2}{5} + \frac{(0)^2}{5} + \frac{(20)^2}{5}\right] - 60$$

$$= [20 + 0 + 80] - 60$$

$$= 100 - 60 = 40$$

$$SSW = TSS - SSB$$

$$= 100 - 40 = 60$$

The various sum of squares (S.S.) along with the degrees of freedom (d.f.) are shown in the following table:

ANOVA Table

Source of variation	Sum of square (S.S.)	Degrees of freedom	Mean Sum of squares (MSS)	F-Radio
Between City	40	3 – 1 = 2	20	$F = \frac{20}{5} = 4$
Within city	60	15 – 3 = 12	5	
Total	100	14		

For $v_1 = 2$ and $v_2 = 12$, the table value of F at 5% I. o.s. = 3·08.

Since the calculated value of F is more than the table value, the null hypothesis is rejected. Hence the population means of the three samples do not seem to be equal.

Example 11 : **The following figures related to the number of units of a product sold in five different areas by four salesmen.**

Area	Number of units			
	A	B	C	D
1	80	100	95	70
2	82	110	90	75
3	88	105	100	82
4	85	115	105	88
5	75	90	80	65

Is there a significant difference in the efficiency of these salesmen ?
(Given that Table value of $F_{.05}$ for $v_1 = 3$, $v_3 = 16$ is 3·24.)

Solution : Let us take the hypothesis that there is no significant difference in the performance of the four salesmen *i.e.* $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

In order to simplify the calculation, subtract 80 from each observation. The deviations and their squares are as follows :

Coded Data

A		B		C		D	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
0	0	20	400	15	225	– 10	100
2	4	30	900	10	100	– 5	25
8	64	25	625	20	400	2	4
5	25	35	1225	25	625	8	64
– 5	25	10	100	0	0	– 15	225
$\Sigma X_1 = 10$	$\Sigma X_1^2 = 118$	$\Sigma X_2 = 120$	$\Sigma X_2^2 = 3250$	$\Sigma X_3 = 70$	$\Sigma X_3^2 = 1350$	$\Sigma X_4 = -20$	$\Sigma X_4^2 = 418$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 10 + 120 + 70 - 20 = 180$$

$$C.F. = \frac{T^2}{N} = \frac{(180)^2}{20} = 1620$$

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + + \Sigma X_3^2 + \Sigma X_4^2\ ] - C.F.$$

$$= [118 + 3250 + 1350 + 418] - 1620$$

$$= 5136 - 1620 = 3516$$

$$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4}\right] - C.F.$$

$$= \left[\frac{(10)^2}{5} + \frac{(120)^2}{5} + \frac{(70)^2}{5} + \frac{(-20)^2}{5}\right] - 1620$$

$$= [20 + 2880 + 980 + 80] - 1620$$

$$= 3960 - 1620 = 2340$$

$$SSW = TSS - SSB$$

$$= 3516 - 2340 = 1176$$

The various sum of squares (S.S.) along with the degrees of freedom (d.f.) are shown in the following table :

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Radio
Between City	2340	4 – 1 = 3	780	$F = \frac{780}{73 \cdot 5} = 10 \cdot 61$
Within city	1176	20 – 4 = 16	73·5	
Total	3516	19		

For $v_1 = 3$ and $v_2 = 16$, the table value of F at 5% I. o.s. = 3·24.

Since, the calculated value of F is greater than the table value, the null hypothesis is rejected. Hence there is a significant difference in the efficiency of the four salesmen.

Example 12. **Complete the following incomplete ANOVA table :**

Source of Variation	Sum of square (S.S)	Degree of Freedom (d.f.)	Mean sum of squares (MSS)	F-test
Between	–	$v_1 = 2$	5	F = ?
Within	14	$v_2 = -$	–	
Total	–	$v = 9$		

Solution : We get $v_2 = v - v_1 = 9 - 2 = 7$

$$\frac{SSB}{v_1} = MSB \quad \Rightarrow \quad SSB = MSB \times v_1 = 5 \times 2 = 10$$

$$TSS = SSB + SSW = 10 + 14 = 24$$

$$SSW = 14, \; MSW = \frac{SSW}{v_2} = \frac{14}{7} = 2$$

$$F = \frac{MSB}{MSW} = \frac{5}{2} = 2 \cdot 5$$

COMPLETE TABLE

Source of Variation	Sum of square (S.S)	Degree of Freedom	Mean sum of squares (MSS)	F-test
Between	10	2	5	$F = \frac{5}{2} = 2 \cdot 5$
Within	14	7	2	
Total	24	9	–	

Example 13 : **A company appoints four salesman A, B, C and D and observes their sales in three seasons in summer, winter and monsoon. The figure (in lakhs) are given in the following tables :**

Seasons	Salesmen			
	A	B	C	D
Summer	36	36	21	35
Winter	28	29	31	32
Monsoon	26	28	29	29

Do the salesman differ significantly in their performance ? (or Carry out an Analysis of Variance).

Solution : The above data are classified according to two criteria : (i) salesmen, and (ii) seasons. It is a two way classification.

Let us take the null hypothesis that there is no difference in the performance of the salesmen. This hypothesis means that there is no difference between the sales of salesmen and off seasons.

In order to simplify the calculation, we subtract 30 from each observation the deveations and their squares are as follows :

	Coded Data					Squares of Coded Data			
Seasons	X_1	X_2	X_3	X_4	Row Total	X_1^2	X_2^2	X_3^2	X_4^2
S	6	6	– 9	5	8	36	36	81	25
W	– 2	– 1	1	2	0	4	1	1	4
M	– 4	– 2	– 1	– 1	– 8	16	4	1	1
Column Total	0	3	– 9	6	T = 0	56	41	83	30

$$T = 0 + 3 - 9 + 6 = 0$$

$$C.F. = \frac{T^2}{N} = \frac{(O)^2}{12} = 0$$

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2] - C.F.$$

$$= 56 + 41 + 83 + 30 - 0 = 210$$

$$SSC = \left[\frac{(0)^2}{3} + \frac{(3)^2}{3} + \frac{(-9)^2}{3} + \frac{(6)^2}{3}\right] - C.F.$$

$$= 0 + 3 + 27 + 12 - 0 = 42$$

$$SSR = \left[\frac{(8)^2}{4} + \frac{(0)^2}{4} + \frac{(8)^2}{4}\right] - C.F. = 16 + 0 + 16 - 0 = 32$$

$$SSE = TSS - SSC - SSR = 210 - 42 - 32 = 136$$

ANOVA Table

Source of Variation	Sum of square (S.S)	Degree of Freedom	Mean sum of squares (MSS)	F-Ratio
Between Columns (Salesmen)	42	3	$\frac{42}{3} = 14$	$F = \frac{22 \cdot 63}{14} = 1 \cdot 62$
Betwen Rows (Seasons)	32	2	$\frac{32}{2} = 16$	$F = \frac{22 \cdot 67}{16} = 1 \cdot 42$
Residual/Error	136	$3 \times 2 = 6$	$\frac{136}{6} = 22 \cdot 67$	
Total	**210**	**11**		

Interpetation

(i) **For salemen :** The calculated value of F = 1·62

Table value of F for (6, 3) d.f. at 5% l.o.s. = 8·94

Since the calculated value of F is less than the table value, we accept the null hypothesis and conclude that the sales of different salesmen do not differ significantly.

(ii) **For salemen :** The calculated value of F = 1·42

Table value of F for (6, 2) d.f. at 5% l.o.s. = 5·15

Since the calculated value of F is less than the table value, we accept the null hypothesis and conclude that there is not significant difference in the seasons so far as sales ae concerned

Example 14 : **To study the performance of three detergents and three different water temperatures, the following 'whiteness' readings were obtained with specially designed equipment :**

Water Temperature	Detergent A	Detergent B	Detergent C
Coled Water	57	55	67
Warm Water	49	52	68
Hot Water	54	46	58

Perform a two-way analysis of variance, using 5 percent level of significance.

Solution : Let us take the null hpothesis that there is no significant difference in the performance of three detergents due to water temperature and vice versa.

In order to simplify calculatins, let us subtract 50 from each figure. The deviations and their squares ae as follows :

Water Temp.	Coded Data X_1	X_2	X_3	Row Total	Squares of Coded Data X_1^2	X_2^2	X_3^2
Cold.	7	5	17	29	49	25	289
Warm	– 1	2	18	19	1	4	324
Hot	4	– 4	8	8	16	16	64
Column Total	10	3	43	T = 56	66	45	677

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10 + 3 + 43 = 56$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(56)^2}{9} = 348 \cdot 44$$

$$\text{TSS} = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - \text{C.F.}$$

$$= [66 + 45 + 677] - 348 \cdot 44$$

$$= 788 - 348 \cdot 44 = 439 \cdot 44$$

$$\text{SSC} = \left[\frac{(10)^2}{3} + \frac{(3)^2}{3} + \frac{(43)^2}{3}\right] - \text{C.F.}$$

$$= 33.3 + 3 + 616 \cdot 33 - 348 \cdot 44 = 304 \cdot 22$$

$$SSR = \left[\frac{(29)^2}{3} + \frac{(19)^2}{3} + \frac{(8)^2}{3}\right] - C.F.$$

$$= 280 \cdot 33 + 120 \cdot 33 + 21 \cdot 33 - 348 \cdot 44 = 73 \cdot 55$$

$$SSE = TSS - SSC - SSR$$

$$= 439 \cdot 56 - 304 \cdot 22 - 73 \cdot 55 = 61 \cdot 79$$

ANOVA Table

Source of Variation	Sum of square (S.S)	Degree of Freedom	Mean sum of squares (MSS)	F-Ratio
Between Columns (Detergents)	304·22	2	152·110	$F = \frac{152 \cdot 110}{15 \cdot 445} = 9 \cdot 85$
Between Rows (Temperatures)	73·55	2	36·775	$F = \frac{36 \cdot 775}{15 \cdot 445} = 1 \cdot 42$
Residual/Error	61·79	4	15·445	
Total	439·56	8		

Interpetation

(i) For Detergents : The calculated value of F = 9·85

Table value of F for (2, 4) d.f. at 5% I.o.s. = 6·94

Since the calculated value of F is greater than the table value, we reject the null hypothesis and conclude that there is significant different in the three varieties of deregents.

(ii) For Temperature : The calculated value of F = 2·38

Table value of F for (2, 4) d.f. at 5% I.o.s. = 6·94

Since the calculated value of F is less than the table value, we accept the null hypothesis and conclude that temperature do not make a significant difference.

Example 15 : Apply F-test on the following data

X_1	25	30	36	38	31	$\Sigma X_1 = 160$
X_2	31	39	38	42	35	$\Sigma X_2 = 185$
X_3	24	30	28	25	28	$\Sigma X_3 = 135$

Given ($F_{2, 12}$, 5 percent = 3·89)

Solution : Null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$ *i.e.* there is no significant difference in the mean of three samples.

In order to simplify the calculation, subtact 30 from each sample values.

Coded Data

Sample 1		Sample 2		Sample 3	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
−5	25	1	1	−6	36
0	0	9	81	0	0
6	9	8	64	−2	4
8	12	12	144	−5	25
1	5	5	25	−2	4
$\Sigma X_1 = 10$ $n_1 = 5$	$\Sigma X_1^2 = 126$	$\Sigma X_2 = 35$ $n_2 = 5$	$\Sigma X_2^2 = 315$	$\Sigma X_3 = -15$	$\Sigma X_3^2 = 69$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10 + 35 - 15 = 30$$

Correction Factor, $$\text{C.F.} = \frac{T^2}{N} = \frac{(30)^2}{15} = 60$$

TSS = Total sum of squares

$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= [126 + 315 + 69] - 60 = 450$$

SSB = Sum of squares between the samples

$$= \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$

$$= \left[\frac{(10)^2}{5} + \frac{(35)^2}{5} + \frac{(-15)^2}{5}\right] - 60 = [20 + 245 + 45] - 60 = 250$$

$$\text{SSW} = \text{TSS} - \text{SSB} = 450 - 250 = 200$$

ANOVA Table

Source of Variation	Sum of squares	Degree of Freedom (d.f.)	Mean sum of squares (MSS)	F-Ratio
Between Samples	250	2	125	$F = \frac{125}{16.667} = 7.5$
Between Samples	200	12	16.667	
Total	**450**	**14**		

For $v_1 = 2, v_2 = 12$, the table value of F at 5% level of significantce is 3·89 Since the calculated value of F is greater than the table value i.e. $F < F_{0.05}$, we reject the null hypothesis and conclude that there is significant difference in the mean of three samples.

QUESTIONS

1. What is analysis of variance technique ? Explain its basic assumptions and uses.
2. (a) Discuss the assumptions of Analysis of Variance test (or techniques)
   (b) Distinguish between one-way and two-way ANOVA technique.
3. Discuss the technique of analysis of variance with an illustration for one-way classification.
4. Describe the technique of ANOVA for two-way classification.
5. What do you understand by analysis of variance? Explain the assumptions in an analysis of variance.
6. What is analysis of variance problem? Comment on the variance between the samples and within the samples.
7. What are the objectives, assumptions and uses of analysis of variance?
8. What is analysis of variance? Mention its applications.
9. Explain the meaning and significance of ANOVA. How is an ANOVA table set up and how a test is performed.

❀❀❀