CHAPTER ELEVEN
TEST OF HYPOTHESIS

11.1. INTRODUCTION

The objective of the statistical analysis is to infer about the population parameter. The inference is drawn on the basis of sample statistic and using some statistical procedure. The first step of the procedure is to assume a value about population parameter. The statistical procedure to verify the assumption on the basis of sample statistic is known as test of hypothesis.

11.2. TERMS RELATED TO TEST OF HYPOTHESIS

Parameter : The unknown characteristic of a population is known as parameter. Thus, if $x \sim N(\mu, \sigma^2)$, μ and σ^2 are two unknown characteristics of the distribution of x. These two unknown characteristics μ and σ^2 are known as parameters.

Statistic : Any function of sample observations is known as sample statistic. For example, let $x_1, x_2, \ldots, x_n$ be n observations of a random sample drawn from a population. The sample mean and sample variance of this sample are given, respectively by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$

Here $\bar{x}$ is a sample statistic, s^2 is also another sample statistic. If the sample is a simple random sample from the population, $\bar{x}$ is an unbiased estimator of population mean and s^2 is an unbiased estimator of population variance.

Hypothesis : Any statement, description or specification about population parameter or population is called hypothesis. The hypothesis is of two types. These are (a) parametric hypothesis, and (b) non-parametric hypothesis. Each hypothesis is again classified into two. These are (a) null hypothesis, and (b) alternative hypothesis.

Parametric Hypothesis : Any statement or assumption regarding parameter of a population is known as parametric hypothesis.

Non-Parametric Hypothesis : The hypothesis which does not specify anything regarding the parameter of the population, rather it specifies other characteristic of the population, is called non-parametric hypothesis.

Null Hypothesis : According to R. A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption it is true. It is denoted by H_o. For example, let $x \sim N(\mu, \sigma^2)$, where μ and σ^2 are parameters. Let us assume that μ has a specified value μ_0 (say). One need to varify whether this assumption is true or false. Then the null hypothesis is

$$H_o : \mu = \mu_0$$

This null hypothesis is related to the parameter of a distribution and hence is called parameteric hypothesis.

Again, let us consider that we have two independent sample observations, where the sample observations are :

First sample : $x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots, x_{1n_1}$

Second sample : $x_{21}, x_{22}, x_{23}, \ldots\ldots\ldots, x_{2n_2}$

One may assume that the samples are drawn from same population. To verify this assumption, let us specify the hypothesis

H_o : Samples are drawn from a population having same median.

This type of hypothesis is called non-parametric hypothesis since it does not specify any value regarding the population parameter.

Alternative Hypothesis : Any statement against the null hypothesis is called alternative hypothesis. It is usually denoted by H_A. Thus, if the null hypothesis regarding population parameter μ is

$$H_o : \mu = \mu_0$$

the alternative hypothesis are

(i) $H_A : \mu \neq \mu_0$

(ii) $H_A : \mu > \mu_0$

(iii) $H_A : \mu < \mu_0$

All these three statements are complementary to the null hypothesis and these are known as alternative hypothesis.

Again, if the assumption regarding a sample is that the sample observations are drawn randomly. This assumption needs to be varified. Thus the null hypothesis is

H_o : the sample is drawn randomly.

The statement against this null hypothesis is

H_A : the sample is not drawn randomly.

Here H_A is the alternative hypothesis.

Test of Hypothesis : The statistical procedure which is used to verify any pre-specified statement or description regarding population or population parameter is known as test of hypothesis.

The statistical procedure also signifies the evidence regarding null hypothesis and hence it is also called test of significance. The evidence is derived using sample statistic.

Simple Hypothesis : The hypothesis which explains all the parameters of the distribution is called simple hypothesis. Thus if $x \sim N(\mu, \sigma^2)$ and the null hypothesis to be tested is

$H_o : \mu = \mu_0$, against $H_A : \mu \neq \mu_0$ is called simple hypothesis if $\sigma^2 = \sigma_0^2$ (say) is also specified. Here one parameter (σ^2) is specified and another one is to be verified according to specification.

Composite Hypothesis : The hypothesis through which all parameters are not specified is known as composite hypothesis. Thus, in case of normal population if

$H_o : \sigma^2 = \sigma_0^2$ (say), against $H_A : \sigma^2 > \sigma_0^2$

without any specification of population mean μ, then H_o is called composite hypothesis.

Test Statistic : The function of sample observations which is used to verify the null hypothesis is called test statistic. Let us consider that we have a random sample of observations $x_1, x_2, \ldots\ldots, x_n$. Assume that this sample is drawn from a normal population $N(\mu, \sigma^2)$. We need to test the hypothesis

$H_o : \mu = \mu_0$, against $H_A : \mu \neq \mu_0$.

We can assume that (i) σ^2 is known and its value is σ_0^2, (ii) σ^2 is unknown.

$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

To test the above hypothesis, the test statistic is

$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ under assumption (ii)

and $Z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}$ under assumption (i)

where $V(\bar{x}) = \sigma^2/n$ and $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Here t and Z both are functions of sample observations and these functions can be used to test the significance of the null hypothesis. Thus 't' and Z are test statistics.

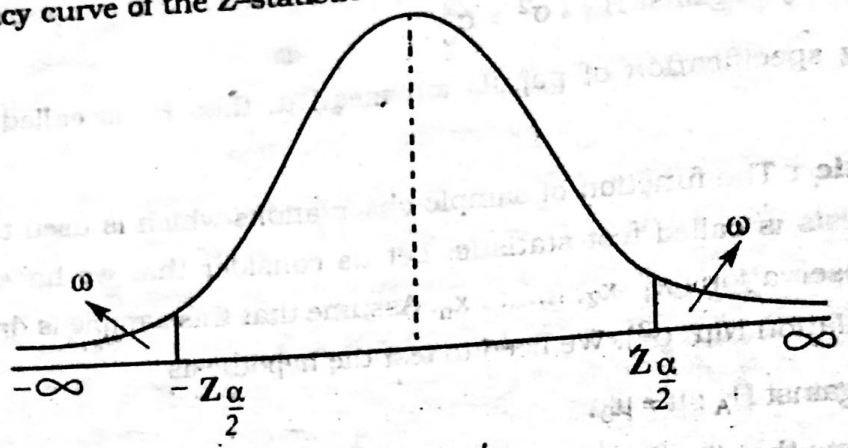
Assumption : The statements regarding population and / or sample which are used to decide the distribution of test statistic are known as assumptions. The test statistic 't' presented above follows Student's 't' distribution under the assumption that the sample observations are independently drawn from a normal population with unknown variance σ^2. On the other hand, the test statistic Z follows N(0, 1) under the assumption that the sample observations are drawn independently from normal population with known variance σ^2.

One Tailed and Two Tailed Tests : Each null hypothesis is opposed by a statement known as alternative hypothesis. Thus, if $H_o : \mu = \mu_0$, the alternative hypothesis may be

(i) $H_A : \mu > \mu_0$, (ii) $H_A : \mu < \mu_0$, (iii) $H_A : \mu \neq \mu_0$.

If alternative hypothesis (i) or (ii) is used in any test, the test is called one-tailed test. Here $H_A : \mu > \mu_0$ is an alternative hypothesis for a right tailed test and $H_A : \mu < \mu_0$ is an alternative hypothesis for a left tailed test. On the other hand, if [illegible] used in any test, the test is called two-tailed test.

Acceptance Region and Critical Region : Each sample statistic follows a specific distribution. The frequency curve of the distribution has a definite shape. The frequency curve of the Z-statistic as discussed above is shown below :



The total area under the curve is 1, i.e.

$$\int_{-\infty}^{\infty} f(z)\, dz = 1$$

The calculated test statistic will fall in any region in the range space. The probability for a space ω is given by

$$\frac{\alpha}{2} = \int_{z_{\alpha/2}}^{\infty} f(z)\, dz \quad \text{or} \quad \frac{\alpha}{2} = \int_{-\infty}^{-z_{\alpha/2}} f(z)\, dz$$

Here α is a very small pre-specified probability. Now, the region in which test statistic falls and leads us to reject the null hypothesis with small probability even null hypothesis is true is known as critical region. Thus, the region ω as shown in the figure is known as critical region and the region bounded by $-Z_{\alpha/2}$ to $Z_{\alpha/2}$ is known as acceptance region. If the test statistic falls within the bounded region bounded by $-Z_{\alpha/2}$ to $Z_{\alpha/2}$, we accept the null hypothesis.

The decision regarding the null hypothesis is made as follows :

(i) If $Z \geq Z_{\alpha/2}$, H_o is rejected in favour of H_A : $\mu > \mu_0$.

(ii) If $-Z \leq -Z_{\alpha/2}$, H_o is rejected in favour of H_A : $\mu < \mu_0$.

(iii) If $|Z| \geq Z_{\alpha/2}$, H_o is rejected in favour of H_A : $\mu \neq \mu_0$.

Level of Significance : The probability level for which the null hypothesis is rejected even it is true is called level of significance. It is denoted by α (unless it is mentioned otherwise, α is taken as 0.05).

Here $P(X \in \omega / H_o) = P(H_o \text{ is rejected}/H_o \text{ is true})$

$= \alpha$

and X is a calculated value of the test statistic. For normal distribution if $\alpha = 0.05$. $Z_{\alpha/2} = 1.96$ if H_A is two-tailed. But $Z_\alpha = 1.64$ for one-sided H_A.

Type-1 Error and Type-11 Error : In taking decision regarding the null hypothesis on the basis of sample statistic, usually two types of error creep in. These error

(i) H_o is rejected when it is true,

(ii) H_o is accepted and H_A is rejected even H_o is not true.

The first error is known as Type–1 error. We have

$P[\text{Type–1 error}] = P(X \in \omega/H_o) = \alpha$

The second error is known as Type–11 error. The probability of Type–11 error is denoted by β. Thus

$P[\text{Type–11 error}] = P(\text{Accept } H_o/H_A \text{ is true}) = \beta$.

Power of Test : The power of the test is $1-\beta$, where β is the probability of Type-11 error. [Prob. of rejecting a false hypothesis is 1-β (P

Degrees of Freedom (d.f) : The number of independent values which are used in deriving a test statistic is known as degrees of freedom (d.f). For example, let $x \sim N(\mu, \sigma^2)$, where $x_1, x_2, \ldots\ldots, x_n$ are independent sample values of x and assume that μ and σ^2 are known. Then

$$Z_i = \frac{x_i - \mu}{\sigma} \sim N(0, 1)$$

$$\sum_{i=1}^{n} Z_i^2 = \frac{\Sigma(x_i - \mu)^2}{\sigma^2} \sim \chi^2 \text{ with n d. f.}$$

Here χ^2 is derived using n independent values of x.

Let us assume that μ is not known and its estimate is $\bar{x}$ and σ^2 known. Then

$$\sum_{i=1}^{n} Z_i^2 = \frac{\Sigma(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2 \text{ with (n–1) d. f.}$$

Here d.f. is the number of independent values of x minus one for one estimated parameter. Thus, d.f is the number of independent values of the variable minus the number of parameter estimated in deriving the test statistic. → c.f from

P–Value : The P-value is the probability of the calculated test statistic under H_o.

Let

$$H_o : \mu = \mu_0 \text{ against } H_A : \mu \neq \mu_0$$

If

$$Z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}$$

is the test statistic for testing the significance of null hypothesis, then

$$P[Z_c] = \int_{Z_c}^{\infty} f(z)\, dz = p$$

where Z_c is the calculated value of Z. This p is called the p–value. If $p \leq \alpha$, H_o is rejected at 100 α% level of significance.

11.3. TEST REGARDING MEAN

Let $x_1, x_2, \ldots\ldots, x_n$ be n sample observations drawn independently from a population with mean μ and variance σ^2. Let us assume that the sample observations follow normal distribution, i.e. $x \sim N(\mu, \sigma^2)$. The problem is to test the null hypothesis

$$H_o : \mu = \mu_0 \text{ against } H_A : \mu \neq \mu_0$$

The assumptions for the test are

(i) σ^2 is unknown and n is small ($n < 30$).

(ii) σ^2 is unknown and n is large ($n \geq 30$).

(iii) σ^2 is known (n is small or large) and its value is, say σ_0^2.

The test statistic to test the significance of H_0 is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \text{ under assumption (i)}$$

This 't' follows Student's 't'-distribution with (n-1) d.f.

The test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0, 1), \text{ under assumption (ii)}$$

The test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1), \text{ under assumption (iii).}$$

where $\bar{x} = \frac{1}{n}\Sigma x_i$, $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$

11.4. TEST OF EQUALITY OF TWO MEANS

Consider $x_{11}, x_{12}, \ldots\ldots, x_{1n_1}$ be n_1 sample observations drawn independently from a normal population $x_1 \sim N(\mu_1, \sigma_1^2)$. Another independent sample observations are $x_{21}, x_{22}, \ldots\ldots, x_{2n_2}$ which are drawn from a population $x_2 \sim N(\mu_2, \sigma_2^2)$. The objective is to test the hypothesis

$$H_0 : \mu_1 = \mu_2, \text{ against } H_A : \mu_1 \neq \mu_2.$$

Assumptions :

(i) Two samples are independent.

(ii) σ_1^2 and σ_2^2 are known values; n_1 and n_2 may be small, may be large.

(iii) $\sigma_1^2 = \sigma_2^2 = \sigma^2$, an unknown value; n_1 and n_2 are small (< 30).

(iv) σ_1^2 and σ_2^2, are unknown values; both n_1 and n_2 are large.

The test statistic to test the significance of null hypothesis is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1), \text{ under assumption (ii).}$$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ under assumption (iii).}$$

where $\bar{x}_1 = \frac{1}{n_1}\Sigma x_{1i}$, $\bar{x}_2 = \frac{1}{n_2}\Sigma x_{2i}$, $s_1^2 = \frac{1}{n_1 - 1}\Sigma(x_{1i} - \bar{x}_1)^2$

$$s_2^2 = \frac{1}{n_2 - 1}\Sigma(x_{2i} - \bar{x}_2)^2 ; \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n}$$

This 't' is distributed as Student's 't' with $(n_1 + n_2 - 2)$ d.f.

The test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1), \text{ under assumption (iv).}$$

Example 11.1 : A random sample of 10 persons are selected and their level of education (in completed years of schooling) is recorded. The sample observations are

x : 5, 2, 0, 4, 16, 14, 10, 10, 6, 5

Do you think that the average schooling years of the persons in population is 5 ?

Solution : Assumption $x \sim NID(\mu, \sigma^2)$, where σ^2 is unknown. The hypothesis to be tested is

$$H_0 : \mu = \mu_0 = 5, \text{ against } H_A : \mu \neq \mu_0$$

The sample mean and sample variance are, respectively

$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{72}{10} = 7.2, \; s^2 = \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] = \frac{1}{9}\left[758 - \frac{(72)^2}{10}\right] = 23.96$$

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.2 - 5}{4.89/\sqrt{10}} = \frac{2.2}{1.55} = 1.42.$$

Comment : $|t| < t_{.05,9} = 2.262$. H_0 is accepted. The average schooling years can be considered as 5.

Example 11.2 : The fasting blood sugar (FBS, x) of 15 randomly selected patients are given below :

110, 118, 130, 140, 142, 146, 112, 100, 95, 98, 96, 122, 123, 124, 130

Do you think that the mean FBS of patients in the population is 110? It is known that the population variance of FBS is $\sigma^2 = 300$. If σ^2 is unknown, how would you conclude?

Solution : Let us assume that the FBS, $x \sim N(\mu, 300)$. We need to test the hypothesis

$$H_0 : \mu = \mu_0 = 110, \text{ against } H_A : \mu \neq \mu_0$$

Here n = 15, but σ^2 is known. The test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \text{ where}$$

$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{1786}{15} = 119.07$$

$$\therefore \quad Z = \frac{119.07 - 110}{\sqrt{300/15}} = 2.03$$

Comment : $|Z| > 1.96$. H_o is rejected. The mean FBS of patients is not 110.

If σ^2 is not known, the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \text{ where}$$

$$s^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] = \frac{1}{14}\left[216682 - \frac{(1786)^2}{15}\right] = 287.78$$

The above 't' follows Student's 't' distribution with (n−1) d.f.

We have

$$t = \frac{119.07 - 110}{\frac{16.96}{\sqrt{15}}} = \frac{9.07}{4.379} = 2.07$$

Comment : $|t| < t_{.05,14} = 2.145$. H_o is accepted. The average FBS of patients may be accepted as 110.

Example 11.3 : The price of rice in different markets in a city are recorded as follows :

Price of rice (taka/kg)

22.50,	21.00,	20.25,	22.00,	21.75,	21.50,	21.00,	21.25,	21.00,	21.75,	22.00,	22.25,
20.00,	20.75,	21.00,	21.50,	21.25,	21.75,	20.75,	20.50,	20.50,	20.00,	21.00,	21.75,
20.25,	21.50,	21.60,	21.50,	20.75,	20.70,	21.25,	21.00,	21.75,	20.50,	20.80,	21.00

The shopkeepers of the markets claim that the average price is 21.00 taka. Is the claim of the shopkeepers true?

Solution : Let x be the price of rice, where $x \sim N(\mu, \sigma^2)$. Assume that the prices are collected from randomly selected shops. We need to test

$$H_o : \mu = \mu_0 = 21.00, \text{ against } H_A : \mu \neq \mu_0$$

Here n = 36 (> 30) and σ^2 is not known. Since n is large, the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0, 1)$$

Here $\bar{x} = \frac{1}{n}\Sigma x = \frac{761.60}{36} = 21.15$

$$s^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] = \frac{1}{35}\left[16125.565 - \frac{(761.60)^2}{36}\right] = 0.3855$$

$$\therefore Z = \frac{21.15 - 21.00}{\frac{0.6209}{6}} = 1.45$$

Comment : $|Z| < 1.96$. H_o is accepted. The shopkeeper's claim is true.

Example 11.4 : The average monthly expenditure of 100 randomly selected students from different private universities is recorded as 12480.00 taka. The variance of the monthly expenditure is 4194236.15. Do you think that the average expenditure of all students of private universities is 10000.00 taka?

Solution : Let x be the monthly expenditure of a student. It is assumed that $x \sim N(\mu, \sigma^2)$. We need to test the null hypothesis

$H_0 : \mu = \mu_0 = 10000.00$, against $H_A : \mu \neq \mu_0$

Since n is large, the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{12480.00 - 10000.00}{\frac{2047.98}{10}} = 12.11$$

Comment : $|Z| > 1.96$, H_0 is rejected. The average expenditure per student cannot be considered as 10000.00 taka.

Example 11.5 : The number of computer scientists coming out from two different universities A and B are employed in different organizations to do job related to computer. The numbers are given for different years as follows :

University	Number of graduates employed in computer related job
A	18, 16, 15, 20, 18, 15, 12
B	20, 14, 12, 22, 16, 14, 15, 10, 12, 18, 10

Do you think that the employment facility for both the universities are similar?

Solution : Let x_1 be the number of graduates of university A and x_2 be the number of graduates of university B. Assume that $x_1 \sim N(\mu_1, \sigma_1^2)$, $x_2 \sim N(\mu_2, \sigma_2^2)$. Also assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We need to test the hypothesis

$H_0 : \mu_1 = \mu_2$, against $H_A : \mu_1 \neq \mu_2$

Since $n_1 = 7 (< 30)$, $n_2 = 11 (< 30)$ and σ^2 is not known, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where}$$

$$\bar{x}_1 = \frac{1}{n_1}\Sigma x_{1i} = \frac{114}{7} = 16.29, \qquad \bar{x}_2 = \frac{1}{n_2}\Sigma x_{2i} = \frac{163}{11} = 14.82$$

$$s_1^2 = \frac{1}{n_1 - 1}\left[\Sigma x_{1i}^2 - \frac{(\Sigma x_{1i})^2}{n_1}\right] \qquad s_2^2 = \frac{1}{n_2 - 1}\left[\Sigma x_{2i}^2 - \frac{(\Sigma x_{2i})^2}{n_2}\right]$$

$$= \frac{1}{6}\left[1898 - \frac{(114)^2}{7}\right] = 6.905 \qquad = \frac{1}{10}\left[2569 - \frac{(163)^2}{11}\right] = 15.364$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{6 \times 6.905 + 10 \times 15.364}{7 + 11 - 2} = 12.192$$

$$s = \sqrt{s^2} = \sqrt{12.192} = 3.49$$

The statistic 't' is distributed as Student's 't' with, $(n_1 + n_2 - 2)$ d.f.

$$\text{Here } t = \frac{16.29 - 14.82}{\sqrt{12.192\left(\frac{1}{7} + \frac{1}{11}\right)}} = \frac{1.47}{1.688} = 0.87$$

Comment : $|t| < t_{.05,16} = 2.12$. H_o is accepted. The employment facility for the students of both universities is similar.

Example 11.6 : The daily temperature (in ^{0}celcius) of two months during summer season are shown below :

Months	Daily temperature (in ^{0}celcius)
1	32, 34, 31, 33, 35, 36, 34, 34, 34, 35, 32, 33, 33, 33, 32, 32, 34, 33, 32, 34, 32, 31, 33, 34, 35, 34, 33, 33, 33, 34, 34
2	34, 34, 35, 35, 35, 35, 35, 35, 36, 37, 34, 33, 34, 35, 34, 34, 36, 34, 33, 34, 32, 33, 34, 36, 35, 35, 35, 34, 35, 34

Do you think that the temperature of both the months are similar?

Solution : Let x_1 be the temperature of month–1 and x_2 be the temperature of month–2. Assume that $x_1 \sim N(\mu_1, \sigma_1^2)$, $x_2 \sim N(\mu_2, \sigma_2^2)$. Also assume that $\sigma_1^2 = \sigma_2^2$. We have $n_1 = 31$, $n_2 = 30$. We need to test the hypothesis

$$H_o : \mu_1 = \mu_2, \text{ against } H_A : \mu_1 \neq \mu_2$$

Since $n_1 = 31$ (> 30), $n_2 = 30$ (≥ 30), the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1).$$

Here

$$\bar{x}_1 = \frac{1}{n_1}\Sigma x_{1i} = \frac{1032}{31} = 33.29, \qquad \bar{x}_2 = \frac{1}{n_2}\Sigma x_{2i} = \frac{1035}{30} = 34.5$$

$$s_1^2 = \frac{1}{n_1 - 1}\left[\Sigma x_{1i}^2 - \frac{(\Sigma x_{1i})^2}{n_1}\right] \qquad s_2^2 = \frac{1}{n_2 - 1}\left[\Sigma x_{2i}^2 - \frac{(\Sigma x_{2i})^2}{n_2}\right]$$

$$= \frac{1}{30}\left[34398 - \frac{(1032)^2}{31}\right] \qquad = \frac{1}{29}\left[35739 - \frac{(1035)^2}{30}\right]$$

$$= 1.41 \qquad = 1.09$$

$$\therefore Z = \frac{33.29 - 34.5}{\sqrt{\frac{1.41}{31} + \frac{1.09}{30}}} = -4.23$$

Comment : $|Z| > 1.96$. H_o is rejected. The temperature in two months are not similar.

Example 11.7 : The average salary of 15 female workers of Industry–1 is 1125.00 taka and standard deviation of salary is 75.00 taka. The average salary of 20 female workers of Industry–2 is 1325.00 taka and standard deviation is 225.00 taka. Are the female workers of two industries similarly paid?

Solution : Let x_1 be the monthly salary of a female worker in Industry–1 and x_2 be the monthly salary of a female worker in Industry–2. Assume that $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$. Also assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say). We need to test the hypothesis

$$H_o : \mu_1 = \mu_2, \text{ against } H_A : \mu_1 \neq \mu_2$$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where}$$

$s^2 = \frac{(n_1-1)\, s_1^2 + (n_2-1)\, s_2^2}{n_1 + n_2 - 2}$. This 't' follows Student's 't' distribution with $(n_1 + n_2 - 2)$ d. f. We have

$s_1^2 = 5625.00,\ s_2^2 = 50625.00,\ n_1 = 15,\ n_2 = 20$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)\, s_2^2}{n_1 + n_2 - 2} = \frac{14 \times 5625.00 + 19 \times 50625.00}{15 + 20 - 2} = 31534.09$$

$$\therefore t = \frac{1125.00 - 1325.00}{\sqrt{31534.00\left(\frac{1}{15} + \frac{1}{20}\right)}} = \frac{-200}{60.65} = -3.30$$

Comment : $|t| > t_{.05,33} = 2.035$, H_o is rejected. The salary of female workers in two industries are not similar.

11.5. TEST OF EQUALITY OF TWO CORRELATED MEANS

Let $x_1, x_2, \ldots\ldots, x_n$ be n observations recorded from n sample points in an occasion. Let $y_1, y_2, \ldots\ldots, y_n$ be another n observations recorded from same n sample points in a different occasion. Here it is assumed that $(x_i, y_i;\ i=1, 2, \ldots, n)$ is the i–th pair of observation from i–th sample point. Let us also assume that $x_i \sim N(\mu_1, \sigma^2)$ and $y_i \sim N(\mu_2, \sigma^2)$. The problem is to test the hypothesis

$$H_o: \mu_1 = \mu_2, \quad \text{against } H_A: \mu_1 \neq \mu_2$$

The hypothesis implies that $H_o: \mu_1 - \mu_2 = 0$. To test this hypothesis, we can consider the differences of pairs of observations, where the difference of i–th pair of observation is

$$d_i = x_i - y_i;\ i = 1, 2, \ldots\ldots, n$$

The variance of d_i is

$$s_d^2 = \frac{1}{n-1}\Sigma(d_i - \bar{d})^2, \text{ where } \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

The test statistic for this hypothesis is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

This 't' follows Student's 't' distribution with (n–1) d.f. Hence, if $|t| \geq t_{.05,n-1}$, H_o is rejected.

Example 11.8 : The following data represent the blood sugar of a group of patients before (B) and after (A) of a specific treatment. Is the treatment successful?

Blood Sugar [(B) x] :	142,	146,	156,	120,	138,	160,	150,	155,	180,	210
Blood Sugar [(A) y] :	120,	130,	120,	115,	100,	160,	150,	138,	160,	180

The sample observations can be arranged as follows :

Sample / Observations	1	2	i	k	Total
1	y_{11}	y_{21}	y_{i1}	y_{k1}	
2	y_{12}	y_{22}	y_{i2}	y_{k2}	
⋮	⋮	⋮	⋮	⋮	
j	y_{1j}	y_{2j}	y_{ij}	y_{kj}	
n_i	y_{1n_1}	y_{2n_2}	y_{in_i}	y_{kn_k}	
Total y_i	y_1	y_2	y_i	y_k	$y_{..} = G$
Mean	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_i$	$\bar{y}_k$	$\bar{y}_{..}$

The total variation of the observations can be measured by a quantity

Total Sum of squares, $TSS = \Sigma\Sigma(y_{ij} - \bar{y}_{..})^2 = \Sigma\Sigma y_{ij}^2 - C.T.$, where

$$C.T = \frac{G^2}{n}, \; n = n_1 + n_2 + \ldots\ldots + n_k$$

If the hypothesis is true, then the sample means should be similar. If not, there should be a variation among the sample means. This variation can be measured by a quantity proportional to

$$\Sigma(\bar{y}_i - \bar{y}_{..})^2$$

Again, since the observations of i-th sample are drawn from i-th population with mean μ_i, all observations are supposed to be similar. In practice, due to some unavoidable circumstances there may be variation within the observations of a sample. This variation can be measured by a quantity proportional to

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Similar variation from all observations can be measured by a quantity proportional to

$$\Sigma\Sigma(y_{ij} - \bar{y}_i)^2$$

Therefore, it is noted that the total variation of the sample observations is due to the two main identified sources of variations, viz. variation between samples and variation within sample. Hence, the total variation of the observations can be partitioned into two pre-identified sources of variation as follows :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}_{..})\right]^2$$

$$= \Sigma\Sigma (y_{ij} - \bar{y}_i)^2 + \Sigma\Sigma (\bar{y}_i - \bar{y}_{..})^2 + 2\,\Sigma\Sigma (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}_{..})$$

$$= \Sigma\Sigma (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}_{..})^2$$

$$\therefore \; TSS = SS(\text{within}) + SS(\text{Between})$$

The above technique of partitioning the total variation of a set of observations into component variations according to pre-identified sources of variation is known as analysis of variance. The technique has first been evolved by R. A. Fisher, the reknowned statistician.

Definition : According to R. A. Fisher, "The analysis of variance is a technique to partition the variance caused by a set of causes in a set of observations into component variations according to pre-identified sources of variation."

The sources of data for analysis of variance is design of experiments, where design of experiments is the technique of data collection through controlled experiment. The different terms related to design of experiment are discussed below:

Experiment : It is the controlled act through which data are collected according to some pre-determined objective. For example, the cultivation of different varieties of rice to select a best variety, use of different medicines on pre-selected patients to identify best medicine for a particular disease.

Treatment : The characteristics of any phenomenon which are under investigation are known as treatment. For example, if out of 5 doses of nitrogen one dose is to be selected as best dose for cultivating rice, the nitrogen fertilizer is the treatment having 5 levels. If the experiment is done to select the computers of best company, then the different producers are the treatments.

Experimental Plot : The experimental material on which a treatment is allotted once is known as experimental plot. For example, in medicine testing experiment a medicine is given to a patient once, then the patient is an experimental plot; a rice variety is cultivated in a plot of land once, the agricultural plot constitutes an experimental plot; a new poultry feed is given to a group of chicks, then the group of chicks constitute the experimental plot, etc.

Experimental Result : The outcome of a treatment from an experimental plot is known as experimental result. Total experimental results from different plots used in experiment are usually under analysis in analysis or variance.

Replication : Each treatment is replicated in several plots in any experiment. The repitition of treatments to different plots is known as replication. It helps the researcher in reducing the error in experimentation and hence increase the precision of the analysis.

Randomization : The method of allocation of treatment to plots by any random technique avoiding the personal bias of the researcher is known as randomization. Randomization is also used to reduce the experimental error and hence it increases efficiency of the experiment.

Local Control : It is also used to reduce the experimental error. The method of selecting homogeneous plots to allocate a treatment in one replication is known as local control. Sometimes the technique is applied to select all homogeneous

plots required for any experiment. For example, in an agricultural experiment using irrigation, all plots must be of same height so that water is spread at same level in all plots.

The replication, randomization and local control are the 3 basic principles of design of experiment. All 3 techniques are used to reduce the experimental error and hence to increase the efficiency of the experiment.

Block : A homogeneous group of experimental plots constitute a block. For example, in a feeding trial experiment in diary project a group of cows of same age or same type constitute a block.

Experimental Error : The experimental result of a treatment is not influenced only by the treatment itself. There are many other causes which affect the outcome of a treatment, where these causes are out of control of the researcher. Such uncontrolled source of causes are known as experimental error. The variation in the outcomes due to uncontrolled sources of factors is known as error variation. For example, in agricultural experiment, the draught or the excess rain may damage the crop. So draught or flood is the external source of variation causing experimental error.

One-way Classification : Let there be k treatments to be investigated in an experiment. These k treatments are to be allocated to n plots such that i–th treatment (i = 1, 2, ... , k) is replicated in n_i plots. The experimental result of j–th replication of i–th treatment is y_{ij}. The variation in y_{ij} observations can be partitioned into two sources of variation as follows :

$$\Sigma\Sigma (y_{ij} - \bar{y}_{..})^2 = \Sigma n_i (\bar{y}_i - \bar{y}_{..})^2 + \Sigma\Sigma (y_{ij} - \bar{y}_i)^2$$

TSS = SS (Treatment) + SS (error) = $S_1 + S_2$

Since the total variation of y_{ij} observations is partitioned into one pre-identified source of variation except error, this analysis of variance is known as one-way classification. The classified information are presented in a table known as analysis of variance table (ANOVA Table) as shown below :

ANOVA Table

Sources of variation	d.f	S.S	$MS = \frac{SS}{df}$	F_{cal}	$F_{.05}$
Treatment	k–1	S_1	$s_1 = \frac{S_1}{k-1}$	s_1/s_2	
Error	n –k	S_2	$s_2 = \frac{S_2}{n-k}$		
Total	n–1				

The above analysis is performed under some assumptions which are mentioned below :

Assumptions in Analysis of Variance :

The model for one-way classification is

$y_{ij} = \mu + \alpha_i + e_{ij}$; $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$

where μ = general mean, α_i = effect of i-th treatment, e_{ij} = random error corresponding to j-th replication of i -th treatment. e_{ij} is normally and independently distributed with mean zero $[E(e_{ij}) = 0]$ and common variance σ^2 $[V(e_{ij}) = \sigma^2]$.

The above assumption is needed to perform F-test as shown in ANOVA Table. If $F_{cal} \geq F_{.05}$ with (k-1) and (n-k) d.f., H_o of equality of means is rejected.

The rejection of H_o of equality of means leads us to test the hypothesis

$$H_o : \mu_i = \mu_i', \text{ against } H_A : \mu_i \neq \mu_i' ; i \neq i' = 1, 2, \ldots, k.$$

The test statistic for this hypothesis is

$$t = \frac{\bar{y}_i - \bar{y}_i'}{\sqrt{MS(error)\left(\frac{1}{n_i} + \frac{1}{n_i'}\right)}}$$

This 't' is distributed as student's 't' with (n-k) [d.f. of error] d.f.

Example 11.10 : In cultivating maize the maize seeds are sown in different plots in rows and columns. The number of seeds germinated in different plots are shown below :

Variety of maize (i)	Plants per plot (y_{ij})	n_i	Total, y_i	Mean, $\bar{y}_i$
1	5, 10, 12, 18, 16,	5,	61	12.20
2	15, 12, 7, 6, 6, 8	6	54s	9.00
3	10, 12, 10, 10,	4	42	10.50
4	15, 16, 15, 14, 15	5	75	15.00
		20 = n	G = 232	

(i) Are the types of seed similar in respect of germination?

(ii) Is there any difference between 3rd and 4th type of seed?

Solution : (i) Given $n_1 = 5$, $n_2 = 6$, $n_3 = 4$, $n_4 = 5$, $n = n_1 + n_2 + n_3 + n_4 = 20$

$G = 232$. Then

$$C.T = \frac{G^2}{n} = \frac{(232)^2}{20} = 2691.20$$

$$TSS = \sum\sum y_{ij}^2 - C.T = 2974.00 - 2691.20 = 282.80$$

$$SS\ (Maize) = \sum \frac{y_i^2}{n_i} - C.T$$

$$= \frac{61^2}{5} + \frac{54^2}{6} + \frac{42^2}{4} + \frac{75^2}{5} - 2691.20 = 105.00$$

$$SS\ (error) = TSS - SS\ (Maize) = 282.80 - 105.00 = 177.80$$

ANOVA Table

Sources of variation	d.f	SS	$MS = \frac{SS}{d.f.}$	F_{cal}	$F_{.05}$
Maize	3	105.00	35.00	3.15	3.24
Error	16	177.80	11.1125		
Total	19				

Here we need to test

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ , where } \mu_i = \text{mean of } i\text{-th maize}$$

against H_A : At least one of the equality does not hold good.

Comment : $F_{cal} < F_{.05\,;\,3,\,16} = 3.24$. H_o is accepted. The maize varieties are of the same type in respect of germination.

(ii) We need to test

$$H_o : \mu_3 = \mu_4 \text{ against } H_A : \mu_3 \neq \mu_4$$

The test statistic is

$$t = \frac{\bar{y}_3 - \bar{y}_4}{\sqrt{MS(error)\left(\frac{1}{n_3} + \frac{1}{n_4}\right)}}$$

This 't' follows Student's 't' with 16 d.f. But this statistic is not calculated, since all the maize varieties are similar [By F-test].

Example 11. 11 : There are 25 computers in an office. These computers are supplied by 5 companies in lots of 5. During working hours each computer fails to work for some times (in hour). The failure time of different computers are shown below :

Company	Failure time (in hour) of computers (y_{ij})					Total y_i	Mean $\bar{y}_i$
A	1.5,	1.0,	0.5,	2.0,	3.0	8.0	1.60
B	1.0,	0.5,	0.2,	0.5,	0.5	2.7	0.54
C	0.0,	2.0,	2.0,	0.0,	0.0	4.0	0.80
D	2.0,	2.5,	1.0,	2.0,	3.0	10.5	2.10
E	3.0,	3.0,	2.0,	2.5,	2.5	13.0	2.60
						38.2	

(i) Are the computers of different companies alike in failure time?

(ii) Is there any difference between computers of company B and company E?

Solution : Let $x \sim N(\mu_1, \sigma^2)$, $y \sim N(\mu_2, \sigma^2)$. We need to test the hypothesis

$$H_o : \mu_1 = \mu_2, \text{ against } H_A : \mu_1 > \mu_2$$

Here

$d_i = x_i - y_i$: 22, 16, 36, 5, 38, 0, 0, 17, 20, 30

$$\bar{d} = \frac{1}{n}\Sigma d_i = \frac{184}{10} = 18.4, \; s_d^2 = \frac{1}{n-1}\left[\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right] = \frac{1}{9}\left[5094 - \frac{(184)^2}{10}\right] = 189.82$$

The test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{18.4}{13.78/\sqrt{10}} = 4.22$$

Comment : $|t| > t_{.05,9} = 1.83$ (for right tailed test). H_o is rejected. The treatment is not effective.

Example 11.9 : In a diary farm two new feeds are introduced for the milking cows so that milk production is increased. The milk production of randomly selected 10 cows are recorded on two different occasions, where feed–1 is given to the cows at first instance and after sometime feed–2 is given. The milk production (in kg) are shown below :

Milk production (in kg)

Feed–1 (x) :	12.8,	30.6,	26.4,	18.7,	20.5,	30.6,	30.5,	28.6,	25.4,	25.0
Feed–2 (y) :	15.0,	30.0,	32.4,	20.6,	18.2,	28.4,	35.6,	32.4,	25.0,	22.0

Are the two feeds same?

Solution : Let $x \sim N(\mu_1, \sigma^2)$, $y \sim N(\mu_2, \sigma^2)$. We need to test the hypothesis

$$H_o : \mu_1 = \mu_2, \text{ against } H_A : \mu_1 \neq \mu_2$$

We have

$d_i = x_i - y_i = -2.2, 0.6, -6, -1.9, 2.3, 2.2, -5.1, -3.8, 0.4, 3.0$

$$\bar{d} = \frac{1}{n}\Sigma d_i = \frac{-10.5}{10} = -1.05, \; s_d^2 = \frac{1}{n-1}\left[\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right] = \frac{1}{9}\left[104.55 - \frac{(-10.5)^2}{10}\right] = 10.39$$

The test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{-1.05}{\frac{3.2233}{\sqrt{10}}} = -1.03$$

Comment : $|t| < t_{.05,9} = 2.262$ (two tailed test). H_o is accepted. The two feeds are similar.

11.6. TEST OF EQUALITY OF SEVERAL MEANS (ANALYSIS OF VARIANCE) :

Let us consider that we have k samples which are independently drawn from k normal populations, where the j–th observation of i–th sample be x_{ij} ($i=1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$) such that $x_{ij} \sim N(\mu_i, \sigma^2)$. The problem is to test the hypothesis

H_o : $\mu_1 = \mu_2 = \ldots = \mu_k$, against

H_A : At least one of the equality does not hold good.

Solution : Here $n_1 = n_2 = n_3 = n_4 = n_5 = 5 = n$, $n_1 + n_2 + n_3 + n_4 + n_5 = 25$

[Or there are k companies, each company supplied n = 5 computers; so there are nk = 5 × 5 = 25 computers.]

$$C.T = \frac{G^2}{nk} = \frac{(38{\cdot}2)^2}{5\times5} = 58{\cdot}3696.$$

$$TSS = \Sigma\,\Sigma y_{ij}^2 - C.T = 85{\cdot}04 - 58{\cdot}3696 = 26{\cdot}6704$$

$$SS\ (company) = \Sigma\frac{y_i^2}{n} - C.T = \frac{366{\cdot}54}{5} - 58{\cdot}3696 = 14{\cdot}9384$$

$$SS\ (error) = TSS - SS\ (company) = 26{\cdot}6704 - 14{\cdot}9384 = 11{\cdot}732$$

ANOVA Table

Sources of variation	df	SS	$MS = \frac{SS}{df}$	F_{cal}	$F_{.05}$
Company	4	14·9384	3·7346	6·37	2·87
Error	20	11·732	0·5866		
Total	24				

Comment : $F_{cal} > F_{.05;\,4,\,24}$ · H_0 is rejected . The computers of different companies are not similar in respect of failure time. Here

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

[The computers of all companies are similar in respect of failure time.]

H_A : At least one of the equality does not hold good.

[The computers of different companies are different in respect of failure time.]

(ii) We need to test the hypothesis

$H_0 : \mu_2 = \mu_5$ against $H_A : \mu_2 \neq \mu_5$

The test statistic is

$$t = \frac{\bar{y}_2 - \bar{y}_5}{\sqrt{\frac{2MS\ (error)}{n}}} = \frac{0{\cdot}54 - 2{\cdot}60}{\sqrt{\frac{2\times0{\cdot}5866}{5}}} = \frac{-2{\cdot}06}{0{\cdot}4844} = -4{\cdot}25$$

Comment : $|t| > t_{.05,\,20} = 2{\cdot}086$. H_0 is rejected. The computers of company B and those of company E are not similar in respect of failure time.

11.7. TEST REGARDING PROPORTION

Convert P to π and Q = (1 − π)

If the variable under study is qualitative in nature, the parameter equivalent to the mean of the variable is population proportion. For example, let $x_1, x_2, \ldots, x_N$ be the values of the variable X observed in N population units, where $x_i = 1$, if i-th unit possesses some characteristic under study, otherwise $x_i = 0$ (The characteristics are, for example, residential status, occupation, colour of hair, colour of skin, family planning adoption behavior, examination result expressed in grade, etc.)

Let

$$\sum_{i=1}^{N} x_i = A \Rightarrow A \text{ of the units possess a particular characteristic.}$$

Then

$\pi = P = \frac{1}{N}\Sigma x_i = \frac{A}{N}$ = proportion of units possessing the characteristic.

Again, let

$\sum_{i=1}^{n} x_i = a \Rightarrow$ a of the sample units possess the characteristic under study.

Then

$p = \frac{1}{n}\Sigma x_i = \frac{a}{n}$ = sample proportion of units possessing the characteristic, where n is the sample size.

The problem is to test the hypothesis

$H_0 : P = P_0$ (a known value), against $H_A : P \neq P_0$

$\pi = \pi_0$ $\quad \pi \neq \pi_0$

The test statistic is

$$\boxed{Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0, 1), \text{ where } Q_0 = 1 - P_0.}$$

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

Comment : If $|Z| \geq 1{\cdot}96$, H_0 is rejected at 5% level of significance.

Sometimes two independent samples are drawn from two populations, where P_1 and P_2 are the proportions of individuals possessing the characteristic under study in population–1 and population–2, respectively, where

$$\pi_1 = P_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_{1i} = \frac{A_1}{N_1}, \quad \pi_2 = P_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} x_{2i} = \frac{A_2}{N_2}$$

The corresponding sample proportions of P_1 and P_2 are, respectively

$$p_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \quad p_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

The problem is to test the hypothesis

$H_0 : P_1 = P_2$ against $H_A : P_1 \neq P_2$. $\quad \pi_1 \neq \pi_2$

$\pi_1 = \pi_2$

The test statistic to test the significance of this hypothesis is

$$Z = \frac{p_1 - p_2}{\sqrt{P_e Q_e\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1), \text{ where } P_e = \frac{a_1 + a_2}{n_1 + n_2} \text{ and } Q_e = 1 - P_e$$

Example 11.12 : A sample of 15 students are selected from a group of 100 students and their grade in S.S.C. examination is recorded as follows :

Students :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Grade :	B	C	A	D	B	C	D	A	B	C	D	B	C	C	D

Do you think that 10% students get grade A?

Solution : Since we are interested in studying the characteristic grade A, the students who got grade A, for them let us assign value 1, otherwise 0. Then we have n = 15, a = 2 and $p = \frac{2}{15} = 0{\cdot}13$.

We need to test the hypothesis

$H_0 : P = P_0 = 0{\cdot}10$ against $H_A : P \neq P_0$

The test statistic is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{0{\cdot}13 - 0{\cdot}10}{\sqrt{\frac{0{\cdot}10 \times 0{\cdot}90}{15}}} = 0{\cdot}39 \qquad \because Q_0 = 1 - P_0 = 1 - 0{\cdot}10 = 0{\cdot}90$$

Comment : $|Z| < 1{\cdot}96$, H_0 is accepted. It can be considered that 10% students got A.

Example 11.13 : A sample of 25 couples provided information that 8 of them adopted family planning. It is assumed that 40% couples in the population adopt family planning. Does the sample conform with the assumption?

Solution : We need to test the hypothesis

$$H_0 : P = P_0 = 0{\cdot}40 \text{ against } H_A : P \neq P_0$$

The test statistic is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}, \text{ where } p = \frac{a}{n} = \frac{8}{25} = 0{\cdot}32$$

$$= \frac{0{\cdot}32 - 0{\cdot}40}{\sqrt{\frac{0{\cdot}40 \times 0{\cdot}60}{25}}} = -0{\cdot}82 \qquad \because Q_0 = 1 - P_0 = 1 - 0{\cdot}40 = 0{\cdot}60$$

Comment : $|Z| < 1{\cdot}96$, H_0 is accepted. The sample information conforms with the assumption.

Example 11·14 : Five per cent workers of an industry are usually injured every year. In a year 50 workers are investigated and found that 10 of them are injured during work in the industry. Does the sample provide similar information about the overall proportion of injured workers?

Solution : We need to test the hypothesis

$$H_0 : P = P_0 = 0{\cdot}05 \text{ against } H_A : P \neq P_0$$

The test statistic is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}, \text{ where } n = 50,\ p = \frac{a}{n} = \frac{10}{50} = 0{\cdot}20,\ Q_0 = 1 - P_0 = 1 - 0{\cdot}05 = 0{\cdot}95$$

$$= \frac{0{\cdot}20 - 0{\cdot}05}{\sqrt{\frac{0{\cdot}05 \times 0{\cdot}95}{50}}} = 4{\cdot}87$$

Comment : $|Z| > 1{\cdot}96$, H_0 is rejected. The sample proportion of injured workers is not similar to the assumed proportion of injured workers.

Example 11.15 : In an industry 100 workers are working, 25 of them are skilled. In another industry there are 18 skilled workers out of 125 workers. Are the skilled workers similar in both industries?

Solution : We need to test the hypothesis

$$H_0 : P_1 = P_2 \text{ against } H_A : P_1 \neq P_2$$

The test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where } p_1 = \frac{a_1}{n_1} = \frac{25}{100} = 0.25,\ p_2 = \frac{a_2}{n_2} = \frac{18}{125} = 0.144$$

$$= \frac{0.25 - 0.144}{\sqrt{0.19 \times 0.81\left(\frac{1}{100} + \frac{1}{125}\right)}} \quad P = \frac{a_1 + a_2}{n_1 + n_2} = \frac{25 + 18}{100 + 125} = 0.19,\ Q = 0.81$$

$$= 2.01$$

Comment : $|Z| > 1.96$, H_0 is rejected. The proportions of skilled workers in two industries are not same.

Example 11.16 : Among 100 literate couples 60 adopted family planning and among 500 illiterate couples 125 adopted family planning. Are the proportions of adopter couples among literate and illiterate couples similar?

Solution : We need to test the hypothesis

$$H_0 : P_1 = P_2 \text{ against } H_A : P_1 \neq P_2$$

We have

$$p_1 = \frac{a_1}{n_1} = \frac{60}{100} = 0.6,\ p_2 = \frac{a_2}{n_2} = \frac{125}{500} = 0.25,\ P = \frac{a_1 + a_2}{n_1 + n_2} = \frac{60 + 125}{100 + 500} = 0.31$$

$$Q = 1 - P = 1 - 0.31 = 0.69$$

The test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.60 - 0.25}{\sqrt{0.31 \times 0.69\left(\frac{1}{100} + \frac{1}{500}\right)}} = 6.91$$

Comment : $|Z| > 1.96$, H_0 is rejected. The proportions of adopter couples among illiterate and literate couples are not similar.

11.8. TEST OF EQUALITY OF SEVERAL PROPORTIONS

Let x_{ij} $(i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i)$ be the j-th observation of i-th sample drawn from i-th population. Consider that the population units in i-th sample are N_i and A_i is the number of population units possessing a characteristic under study. Then the proportion of units in the population possessing the characteristic is

$$\pi_i = P_i = \frac{A_i}{N_i} \ ; \qquad i = 1, 2, \ldots, k$$

We need to test the hypothesis

$H_0 : P_1 = P_2 = \cdots = P_k$ against H_A : at least one of the equality does not hold good.

$\pi_1 = \pi_2 = \ldots = \pi_k$

Consider that O_i is the number of observation in i-th sample possessing the characteristic, where the sample size is n_i $(i = 1, 2, \ldots, k)$. We have, the sample proportion

$$p_i = \frac{O_i}{n_i} ;\ i = 1, 2, \ldots, k ;\ n = \sum_{i=1}^{k} n_i$$

If the hypothesis is true, then the observations possessing the characteristic in all samples should be same and these are

$$E_i = \frac{n}{k}, \text{ where } n = \sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$$

n = total from just like $N = \sum f_i$

The test statistic for the above hypothesis is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum \frac{O_i^2}{E_i} - n$$

[Note $\sum \frac{O_i^2 - 2O_iE_i + E_i^2}{E_i} = \sum \frac{O_i^2}{E_i} - 2n + n$

This χ^2 follows χ^2 distribution with (k–1) d.f. The χ^2–statistic provides satisfactory information if $n \geq 10$, $k \geq 3$, $\frac{n^2}{k} \geq 10$ and $E_i > 0{\cdot}25$; otherwise if $E_i < 5$, it is to be added to the preceding or following E_i [Slakter (1973), Koehler and Larntz (1980)]. The d.f of χ^2 is $(k_1 - 1)$, where k_1 is the number of samples after merging any sample to the nearest sample or class.

Example 11.17 : The number of female workers in different garments industries are shown below :

Industry :	1	2	3	4	5	Total
No. of female workers, O_i :	125	75	150	55	200	605

Are the proportions of female workers similar in all industries?

Solution : Let P_i be the proportion of female workers in i-th industry, i = 1, 2, 3, 4, 5. We need test the hypothesis

$$H_0 : P_1 = P_2 = P_3 = P_4 = P_5 \text{ ; against } H_A : P_1 \neq P_2 \neq P_3 \neq P_4 \neq P_5.$$

Under H_0, We have

$$E_i : \frac{n}{k} = \frac{605}{5} = 121$$

Therefore,

$$\chi^2 = \sum \frac{O_i^2}{E_i} - n = \frac{1}{121}(125^2 + 75^2 + 150^2 + 55^2 + 200^2) - 605 = 112{\cdot}15$$

Comment : $\chi^2 = 112{\cdot}15 > \chi^2_{.05,4} = 9{\cdot}488$. H_0 is rejected. The proportions of female workers are not similar in all industries.

Example 11.18 : The number of road accidents in 4 different highways in a month are as follows :

Highways :	1	2	3	4	Total
No. of accidents O_i :	50	42	32	82	206

Are the proportions of accidents in all highways same?

Solution : Let P_i be the proportion of accidents in i-th highway, i = 1, 2, 3, 4. We need to test the hypothesis

$H_0 : P_1 = P_2 = P_3 = P_4$ against H_A : At least one equality does not hold good.

Under H_0, $E_i = \frac{n}{k} = \frac{206}{4} = 51{\cdot}5$.

The test statistic is

$$\chi^2 = \sum \frac{O_i^2}{E_i} - n = \frac{1}{51{\cdot}5}(50^2 + 42^2 + 32^2 + 82^2) - 206 = 27{\cdot}24$$

Comment : $\chi^2 = 27{\cdot}24 > \chi^2_{0{\cdot}05,3} = 7{\cdot}815$, H_0 is rejected. The proportions of accidents in highways are not same.

Example 11.19 : In an school final examination the grades of 100 students are as follows :

Grade :	A	B	C	D	E	Total
No. of students, O_i :	10	25	42	13	10	100

Do you think that the grades are in the ratio

A:B:C:D:E::1:2:3:1:1?

Solution : We need to test the hypothesis

$$H_0 : P_1 : P_2 : P_3 : P_4 : P_5 :: 1:2:3:1:1$$

or, $H_0 : 3P_1 = \frac{2}{3} P_2 = P_3 = 3P_4 = 3P_5$, against $H_A : 3P_1 \neq \frac{2}{3} P_2 \neq P_3 \neq 3P_4 \neq 3P_5$

Under H_0, the expected number of students in different grades are

$$E_1 = \frac{100}{8} = 12{\cdot}50,\ E_2 = \frac{2 \times 100}{8} = 25{\cdot}00,\ E_3 = \frac{3 \times \cancel{200}\ 100}{8} = 37{\cdot}50,\ E_4 = E_5 = \frac{\cancel{200}\ 100}{8} = 12{\cdot}50$$

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n = \frac{10^2}{12{\cdot}50} + \frac{25^2}{25{\cdot}00} + \frac{42^2}{37{\cdot}50} + \frac{13^2}{12{\cdot}50} + \frac{10^2}{12{\cdot}50} - 100 = 1{\cdot}56$$

Comment : $\chi^2 = 1{\cdot}56 < \chi^2_{0.05,4} = 9{\cdot}488$. H_0 is accepted. The proportions of students in different grades are in the given ratio.

Example 11.20 : The number of literate persons in 4 different villages are as follows :

Village :	1	2	3	4	Total
No. of literate persons, O_i :	250	450	150	150	1000

Do you think that the proportions of literate persons are similar in all villages.

Solution : Let P_i be the proportion of literate persons in i-th village, i = 1, 2, 3, 4. We need to test the hypothesis

$H_0 : P_1 = P_2 = P_3 = P_4$ against H_A : At least one of the equality does not hold good.

Under H_0, we have $E_i = \frac{n}{k} = \frac{1000}{4} = 250$

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n = \frac{1}{250}(250^2 + 450^2 + 150^2 + 150^2) - 1000$$

$$= 240{\cdot}00$$

Comment : $\chi^2 = 240{\cdot}00 > \chi^2_{.05,3} = 7{\cdot}815$. H_0 is rejected. The proportions of literate persons in different villages are not similar.

11.9. TEST REGARDING VARIANCE

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from $N(\mu, \sigma^2)$, where μ is unknown. The objective is to test the hypothesis

$H_0 : \sigma^2 = \sigma_0^2$ (say) against (i) $H_A : \sigma^2 \neq \sigma_0^2$

or, (ii) $H_A : \sigma^2 > \sigma_0^2$ or (iii) $H_A : \sigma^2 < \sigma_0^2$

Since μ is unknown, it is estimated by $\bar{x}$ and the estimate of σ^2 is

Then the test statistic is

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}$$

This χ^2 is distributed as chi-square with (n–1) d.f.

Comment : Alternative hypothesis (i) $H_A : \sigma^2 \neq \sigma_0^2$. It is a two-sided alternative. Hence, if

$$\chi^2_{cal} \leq \chi^2_{1-\frac{\alpha}{2}, n-1} \text{ or } \chi^2_{cal} \geq \chi^2_{\frac{\alpha}{2}, n-1}$$

H_0 is rejected.

Alternative hypothesis (ii) : $H_A : \sigma^2 > \sigma_0^2$

It is a right sided test and the conclusion is made as follows.

$\chi^2_{cal} \geq \chi^2_{\alpha, n-1}$, H_0 is rejected.

Alternative hypothesis (iii) $H_A : \sigma^2 < \sigma_0^2$

It is a left sided test and the conclusion is made as follows :

$\chi^2_{cal} \leq \chi^2_{1-\alpha, n-1}$, H_0 is rejected.

Again, let us consider that $x_{11}, x_{12}, \ldots, x_{1n_1}$ be a random sample from N (μ_1, σ_1^2) and $x_{21}, x_{22}, \ldots, x_{2n_2}$ be another random sample from N (μ_2, σ_2^2). The two samples are independent and μ_1 and μ_2 are unknown. The estimate of μ_1 is $\bar{x}_1 = \frac{1}{n_1}\Sigma x_{1i}$ and the estimate of μ_2 is $\bar{x}_2 = \frac{1}{n_2}\Sigma x_{2i}$.

The objective is to test the hypothesis

$H_0 : \sigma_1^2 = \sigma_2^2$ against $H_A : \sigma_1^2 \neq \sigma_2^2$

The test statistic is

$$F = \frac{s_1^2}{s_2^2} \text{ , where}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 \text{ and } s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 ; \; s_1^2 > s_2^2$$

This F is distributed as variance ratio with $(n_1 - 1)$ and $(n_2 - 1)$ d.f.

Comment : If calculated $F \geq F_{.05 : n_1-1, n_2-1}$, H_0 is rejected at 5% level of significance provided $\frac{\sigma_1^2}{\sigma_2^2} \geq \sigma_0^2$ (a specified value.)

Example 11.21 : The following observations represent the systolic blood pressure (x, mm of Hg) of some patients :

x : 70, 85, 92, 90, 95, 79, 80, 85, 90, 85, 95.

Do you think that the sample is drawn from a population N (μ, 50)

Solution : We need to test the hypothesis

$H_0 : \sigma^2 = \sigma_0^2 = 50$ against $H_A : \sigma^2 > \sigma_0^2$

We have

$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{946}{11} = 86,\ s^2 = \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] = \frac{1}{10}\left[81930 - \frac{(946)^2}{11}\right] = 57{\cdot}4$$

The test statistic is

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma_0^2} = \frac{(11-1)\,57{\cdot}4}{50} = 11{\cdot}48$$

Comment : $\chi^2 < \chi^2_{.05,\,10} = 18{\cdot}307$. H_0 is accepted.

Example 11.22 : The number of telephone calls received by the emergency word during office hour in different days are as follows :

160, 172, 121, 144, 100, 108, 175, 200, 105, 95, 102

Do you think that the sample follows N (μ, 1500)

Solution : We need to test the hypothesis

$H_0 : \sigma^2 = \sigma_0^2 = 1500$ against $H_A : \sigma^2 < \sigma_0^2$

We have

$$s^2 = \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] = \frac{1}{n-1}\left[213304 - \frac{(1482)^2}{11}\right] = 1363{\cdot}82$$

The test statistic is

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma_0^2} = \frac{(11-1)\,1363{\cdot}82}{1500} = 9{\cdot}09$$

Comment : $\chi^2 > \chi^2_{0.95,10} = 3{\cdot}94$. H_0 is ~~accepted~~. ~~rejected~~ accepted.

Example 11.23 : The number of students admitted in two private universities in different years are as follows :

University–1, x_{1i}	:	155,	165,	170,	190,	220,	250,	250
University–2, x_{2i}	:	300,	355,	360,	360,	360,	400,	400, 400

Are the variations in admission of students in different years in two universities are same?

Solution : Let $x_{1i} \sim N(\mu_1, \sigma_1^2)$, $x_{2i} \sim N(\mu_2, \sigma_2^2)$

We need to test the hypothesis

$H_0 : \sigma_1^2 = \sigma_2^2$ against $H_A : \sigma_1^2 \neq \sigma_2^2$

We have

$$s_1^2 = \frac{1}{n_1 - 1}\left[\Sigma x_{1i}^2 - \frac{(\Sigma x_{1i})^2}{n_1}\right] = \frac{1}{7-1}\left[289650 - \frac{(1400)^2}{7}\right] = 1608{\cdot}33$$

$$s_2^2 = \frac{1}{n_2 - 1}\left[\Sigma x_{2i}^2 - \frac{(\Sigma x_{2i})^2}{n_2}\right] = \frac{1}{8-1}\left[1084825 - \frac{(2935)^2}{8}\right] = 1149{\cdot}55$$

Since $s_1^2 > s_2^2$, the test statistic is

$$F = \frac{s_2^2}{s_1^2} = \frac{1149{\cdot}55}{1608{\cdot}33} = 0{\cdot}71$$

$F = \frac{S_1^2}{S_2^2} = \frac{1608}{1149} = 1{\cdot}399$

Comment : $F < F_{.05,\,7,\,6} = 3{\cdot}51$. H_0 is accepted.

$F_{\alpha/2,\,6,\,7} = 3{\cdot}87$

11.10. TEST OF INDEPENDENCE

Let us consider that a researcher has n units in a sample. The sample observations are classified according to qualitative characters, say A and B as follows :

A \ B	B	Not B	Total
A	a	b	a+b
Not A	c	d	c+d
Total	a+c	b+d	n

Here a, b, c and d are observed frequencies in different cells. The problem is to test the hypothesis

H_0 : The two attributes A and B are independent.

H_A : The two attributes A and B are not independent

Assumption : The marginal totals [a + b, c + d, a + c, b + d] and grand total [n] are fixed.

Under H_0, the expected frequencies corresponding to a, b, c and d are

$$a' = \frac{(a+b)(a+c)}{n}, \qquad b' = \frac{(b+d)(a+b)}{n}$$

$$c' = \frac{(a+c)(c+d)}{n} \text{ and } d' = \frac{(b+d)(c+d)}{n}$$

However, in the above (2 × 2) contingency table if one expected frequency is found out, other expected frequencies are found out automatically since marginal totals are fixed.

The test statistic for the hypothesis is $\chi^2 = \sum\sum \frac{O_{ij}^2}{E_{ij}} - n = \frac{a^2+b^2+c^2+d^2}{a'+b'+c'+d'} - n$

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

This χ^2 is distributed as chi-square with 1 d. f. Hence, [(2−1) × (2−1) d.f.]
if $\chi^2_{cal} \geq \chi^2_{.05,1} = 3{\cdot}84$, H_0 is rejected. Rashid (298)

For large n and for no expected frequency smaller enough the χ^2 statistic is approximated well to chi-square distribution. The good result using above χ^2 statistic is also observed if n ≥ 10, expected frequency is greater than or equal to 0·25 [Slakter (1973)]. The widely used technique for better approximated chi-square in case of smaller sample size (n < 40) is the Yate's correction, where the corrected χ^2 is

$$\chi^2 = \frac{n\left[\,|ad-be| - \frac{n}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

The above χ^2 statistic is used to test the independence of two attributes, where each attribute has two levels. In practice, each attribute may has different levels. Consider that A has r levels and B has C levels. The number of observations of

i-th level of A corresponding to j-th level of B is, say O_{ij}, where the distribution of O_{ij} is shown below :

A \ B	B_1	B_2	...	B_j	...	B_c	Total
A_1	O_{11}	O_{12}		O_{1j}		O_{1c}	R_1
A_2	O_{21}	O_{22}		O_{2j}		O_{2c}	R_2
⋮	⋮	⋮		⋮		⋮	⋮
A_i	O_{i1}	O_{i2}		O_{ij}		O_{ic}	R_i
⋮	⋮	⋮		⋮		⋮	⋮
A_r	O_{r1}	O_{r2}		O_{rj}		O_{rc}	R_r
Total	C_1	C_2		C_j		C_C	n

Under the null hypothesis, the proportions of observation in any cell are equal and are given, for i-th row, by

$$\frac{O_{i1}}{C_1} = \frac{O_{i2}}{C_2} = \dots = \frac{O_{ij}}{C_j} = \dots = \frac{O_{ic}}{C_c} = \frac{R_i}{n}$$

This implies that

$$O_{ij} = \frac{C_j \times R_i}{n} \; ; \; i = 1, 2, \dots r \; ; \; j = 1, 2, \dots, c.$$

This value of O_{ij} is under null hypothesis and is called expected frequency corresponding to i-th row and j-th column and is denoted by E_{ij}. Thus

$$E_{ij} = \frac{R_i \times C_j}{n} \; ; \; i = 1, 2, \dots, r \; ; \; j = 1, 2, \dots, c.$$

Since marginal totals and grand total are fixed, we can find out (r−1) (c−1) E_{ij} values independently . Under H_0, the deviation between O_{ij} and E_{ij} should be minimum. Thus, this deviation can be used as a test statistic, where the test statistic is

$$\chi^2 = \Sigma\Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \Sigma\Sigma \frac{O_{ij}^2}{E_{ij}} - n$$

This χ^2 is distributed as chi-square with (r−1) (c−1) d. f. provided no $E_{ij} < 5$. The widely used technique is that, in case of any $E_{ij} < 5$, it is added to its preceeding or following cell. The d. f. of χ^2 in such a situation is $(r_1 - 1)(c_1 - 1)$, where r_1 and c_1 are the adjusted number of rows and columns, respectively. However, Koehler and Larntz (1980) mentioned that the χ^2-statistic provides satisfactory result if $n \geq 10$, $k \geq 3$ (k is the number of classes), $\frac{n^2}{k} \geq 10$ and $E_{ij} > 0{\cdot}25$.

Example 11.24 : One hundred fifty computer graduates are interviewed and are classified according to their result and job satisfaction. The number of graduates in different classes are shown below :

Result	Job satisfaction		Total
	Yes	No	
Good	22	58	80
Not good	20	50	70
Total	42	108	150

Do you think that the graduates with good result are satisfied with their job?

Solution : The null hypothesis to be tested is

H_0 : Job satisfaction does not depend on good result

against H_A : The good result and job satisfaction are associated.

The test statistic is

$$\chi^2 = \frac{n\,(ad - bc)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} = \frac{150\,(22 \times 50 - 20 \times 58)^2}{80 \times 70 \times 42 \times 108} = 0{\cdot}02$$

Comment : $\chi^2 < \chi^2_{.05,\,1} = 3{\cdot}84$. H_0 is accepted. Therefore, job satisfaction does not depend on good result.

Example 11.25 : Thirty blood pressure patients are classified as follows :

Heart problem	Blood pressure		Total
	High	Not high	
Yes	14	1	15
No	6	9	15
Total	20	10	30

Is there any reason to believe that high blood pressure causes heart problem?

Solution : The hypothesis to be tested is

H_0 : High blood pressure is not the cause of heart problem.

against H_A : Heart problem causes due to blood pressure.

The test statistic is

$$\chi^2 = \frac{n\left[\,|ad - bc| - \frac{n}{2}\right]^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} \quad \because n < 40$$

$$= \frac{30\left[\,|126 - 60| - \frac{30}{2}\right]^2}{15 \times 15 \times 20 \times 10} = 7{\cdot}35$$

Comment : $\chi^2 > \chi^2_{.05,\,1} = 3{\cdot}84$. H_0 is rejected. Heart problem arises due to high blood pressure.

Example 11.26 : A group of couples of child-bearing ages are classified according to wives' education and family planning adoption as follows :

Family planning adoption	Level of education				
	Illiterate	Primary	Secondary	Higher	Total
Yes	8	22	25	20	75
No	17	73	25	10	125
Total	25	95	50	30	200

Do you think that family planning adoption depends on level of education?

Solution : The hypothesis to be tested is

H_0 : Family planning adoption does not depend on level of education.

against H_A : Family planning adoption and level of education are associated.

The expected frequencies are calculated and shown below:

Family planning adoption	Level of education				
	Illiterate	Primary	Secondary	Higher	Total
Yes	9·375	35·625	18·75	11·25	75
No	15·625	59·375	31·25	18·75	125
Total	25	95	50	30	200

The test statistic is

$$\chi^2 = \Sigma\Sigma \frac{O_{ij}^2}{E_{ij}} - n$$

$$= \frac{8^2}{9{\cdot}375} + \frac{22^2}{35{\cdot}625} + \frac{25^2}{18.75} + \frac{20^2}{11{\cdot}25} + \frac{17^2}{15{\cdot}625} + \frac{73^2}{59{\cdot}375} + \frac{25^2}{31{\cdot}25} + \frac{10^2}{18{\cdot}75} - 200$$

$$= 22{\cdot}88$$

Comment : $\chi^2 \geq \chi^2_{0.05,\,3} = 7{\cdot}815$. H_0 is rejected. With the increase in level of education the family planning adoption also increases.

11.11. TEST RELATED TO REGRESSION COEFFICIENT AND CORRELATION COEFFICIENT

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of values recorded from n sample points on two variables X and Y. It is assumed that the variable Y depends on variable X and the linear relationship of Y and X is

$$y = \alpha + \beta x + e, \text{ where}$$

β is the regression coefficient of y on x, α is also a regression parameter and e is the random component. It is also assumed that $e \sim N(0, \sigma^2)$. The population correlation coefficient of X and Y is, say P.

The problem is to test the hypothesis

(i) $H_0 : \beta = 0$ against $H_A : \beta \neq 0$

(ii) $H_0 : P = 0$ against $H_A : P \neq 0$.

The test statistic for hypothesis (i) is

$$t = \frac{b}{\sqrt{\frac{s^2}{SS(x)}}}, \text{ where}$$

$$s^2 = \frac{1}{n-2}[SS(y) - b\,SP(xy)], \quad SS(y) = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$SP(xy) = \Sigma xy - \frac{\Sigma x\,\Sigma y}{n}, \quad b = \frac{SP(xy)}{SS(x)}, \quad SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

This 't' follows Student's 't' with (n–2) d.f.

The test statistic for hypothesis (ii) is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \text{ where } r = \frac{SP(xy)}{\sqrt{SS(x)\ SS(y)}}$$

This 't' is also distributed as Student's 't' with (n–2) d.f.

Comment : $|t| \geq t_{.05,\, n-2}$ leads us to reject the null hypothesis at 5% level of significance.

Example 11.27 : The following are the heights (x in mm) and weights (y in gm) of some animals.

x:	155	158	155	152	159	158	154	153	152	154
y:	65	62	56	56	60	58	58	55	56	60

(i) Is the weight of animals significantly increased due to the increase in height?

(ii) Test the significance of correlation of height and weight.

Solution : We have $n = 10$

$$SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 240308 - \frac{(1550)^2}{10} = 58{\cdot}00$$

$$SS(y) = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 34430 - \frac{(586)^2}{10} = 90{\cdot}40$$

$$SP(xy) = \Sigma xy - \frac{\Sigma x\, \Sigma y}{n} = 90866 - \frac{1550 \times 586}{10} = 36{\cdot}00$$

$$b = \frac{SP(xy)}{SS(x)} = \frac{36{\cdot}00}{58{\cdot}00} = 0{\cdot}62$$

We need to test the hypothesis

$H_0 : \beta = 0$ against $H_A : \beta \neq 0$

The test statistic is

$$t = \frac{b}{\sqrt{\frac{s^2}{SS(x)}}}, \qquad \text{where } s^2 = \frac{1}{n-2}\,[SS(y) - b\,SP(xy)]$$

$$= \frac{0{\cdot}62}{\sqrt{\frac{8{\cdot}51}{58{\cdot}00}}} \qquad = \frac{1}{10-2}\,[90{\cdot}40 - 0{\cdot}62 \times 36] = 8.5$$

$$= 1{\cdot}62 \qquad = 8{\cdot}51$$

Comment : $|t| < t_{.05,\, 8} = 2{\cdot}306$. H_0 is accepted. The data do not provide evidence that weight increases with the increase in height.

(ii) $$r = \frac{SP(xy)}{\sqrt{SS(x)\, SS(y)}} = \frac{36{\cdot}00}{\sqrt{58{\cdot}00 \times 90{\cdot}40}} = 0.50$$

We need to test the hypothesis

$H_0 : P = 0$ against $H_A : P \neq 0$

The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{\cdot}50\sqrt{10-2}}{\sqrt{1-(0{\cdot}50)^2}} = 1{\cdot}63$$

Comment : $|t| < t_{.05,\, 8} = 2{\cdot}306$. H_0 is accepted. The data say that height and weight are not correlated.

EXERCISE

1. (a) What is meant by test of hypothesis? Define null hypothesis, level of significance, critical region with example.

   (b) The following are the per acre production of rice of some farmers:
   Production of rice (md/acre)
   60·6, 70·2, 50·8, 45·7, 80·2, 60·0, 66·0, 72·5, 48·7, 60·3, 70·0
   Is the production data follows a distribution with mean 65·0?

2. (a) What is the difference between t-test and z-test? Explain the terms related to test of hypothesis.

   (b) Two independent random samples are drawn from two populations : The sample observations are

   x: 10·2, 10·6 11·7 15·2 9·6 16·4, 8·7 12·3 14·4

   y: 11·6, 16·7, 17·2, 15·0, 18.7 14·2, 12·7, 11·4, 9·6, 11·0, 13·4

   Do you think that the populations are same?

3. (a) What is meant by contingency table? The following data are related to a sample observations classified into two classes as follows :

	A	Not A	Total
B	40	20	60
Not B	30	10	40
	70	30	100

   Are A and B independent?

   (b) The blood sugar level of a group of patients after and before treatment are as follows :

Before treatment :	92	96	110	120	140	160	210	250	160	180
After treatment :	100	80	100	130	160	120	160	240	100	120

   Is the treatment effective?

4. (a) What is analysis of variance? Define treatment, randomisation, replication, local control with example.

   (b) In a diary farm 4 new feeds are given to different cows of same lactation period. The milk production (kg/day) per cow are given below after introducing the new feed plan in the farm.

New feed	Milk production (kg/day)					
F_1	30·5,	25·5,	32·6,	35·0		
F_2	28·2,	20·5,	26·7,	24·2,	24·0	
F_3	31·2,	32·4,	34·3,	35·0,	37·0	
F_4	18·2,	20·4,	21·2,	19·0,	20·6,	26·2

   (i) Is there any difference among the feeds?

   (ii) Are F_2 and F_3 similar?

5. (a) What is meant by test of hypothesis? What is 't'-test? Write down the uses of t-test.

(b) From a population a random sample of 12 observations are drawn. The sample observations are

10, 12, 11, 9, 15, 17, 10, 10, 14, 16, 18, 12

Do you think that the sample is drawn from a population with variance $\sigma^2 = 45$?

6. (a) Define test statistic, critical region, level of significance, null hypothesis with examples.

(b) The following information are given from two samples :

Sample – 1

$n_1 = 32$, $\Sigma x = 48$, $\Sigma x^2 = 126$

Sample – 2

$n_2 = 40$, $\Sigma y = 100$, $\Sigma y^2 = 3000$

Are the two samples drawn from same population?

7. (a) What is meant by test of significance? Write down the procedure to test the significance of regression coefficient.

(b) The per acre production of jute (y, md) and the per acre cost of cultivation (x, taka) of different farmers are given below :

x:	500·00	650·00	1250·00	850·00	950·00	1000·00	500·00	1200·00
y:	10·50	15·00	18·00	15·00	12·50	14·00	6·00	16·00

Is the per acre cost effective to produce more jute?

8. (a) What is meant by test of independence? What are the assumptions required for test of independenc?

(b) Two companies supply computer in an office with assurance of good service for 2 years. But, in practice, the assurance is not effective. The computers are classified as follows according to the failure of assurance.

Assurance	Company		Total
Satisfaction	A	B	
Yes	15	20	35
No	10	5	15
Total	25	25	50

Do you think that the assurance is fruitful for a certain company?

9. (a) Define parameter, statistic, critical region, level of significance with examples.

(b) What is the need of test of proportion? In a locality 80% people are involved in farming and in another locality the percentage of people involved in farming is 70%. The number of people in former and latter localities are 500 and 1000. Are the percentages of people involved in farming in two localities similar?

10. (a) Define treatment, randomisation, replication, local control and analysis of variance.

(b) In an agricultural farm 5 new varieties of rice are introduced to select a best variety in respect of production. These rice varieties are cultivated in different plots of homogeneous size and shape and in respect of fertility. The production of rice (in kg) per plot are given below :

Rice variety	Production of rice (kg/plot)					
R_1	10·2,	18·6,	15·7,	13·4,	10·2,	11·6,
R_2	12·4,	16·2,	15·0,	15·0,	15·0,	16·2,
R_3	15·5,	18·7,	18·0,	16·0,	14·2,	15·6,
R_4	10·4,	9·6,	10·2,	11·0,	10·8,	10·2,
R_5	11·6,	11·0,	11·4,	12·0,	12·5,	12·0,

(i) Are the rice varieties homogeneous?

(ii) Is there any difference between R_3 and R_4 ?

11. (a) Explain test of independence with example. Write down the assumptions for test of independence.

(b) The proportion of people above 60 years is assumed to be 0·15 in a locality. A sample of 25 people shows that there are 2 people above 60 years. Is there any difference in the assumed and observed proportions?

12. (a) Write down the use of t-test and z-test. Define null hypothesis, alternative hypothesis, critical region, level of significance with example.

(b) The following information are given from two independent samples drawn from two populations.

Sample–1

$n_1 = 8,\ \Sigma x = 16,\ \Sigma x^2 = 40.$

Sample–2

$n_2 = 12,\ \Sigma y = 48,\ \Sigma y^2 = 220.$

Can you say that the two populations are same?