Unit-IV	**Small Sample Tests, Chi-Square Tests & F- tests**
4.1	Small Sample Tests / **Student's 't' test**
4.2	Test for Single Mean
4.3	Test for Difference of Means • Independent samples • Paired samples
4.4	Chi-Square Test • Test for Independence of Attributes • Goodness of Fit
4.5	F- test for Equality of Two Variances

4.1 Small Sample Tests - Student's 't' test

In the previous chapter we have discussed problems relating to large samples. The large sampling theory is based upon two important assumptions such as

(a) The random sampling distribution of a statistic is approximately normal and

(b) The values given by the sample data are sufficiently close to the population values and can be used in their place for the calculation of the standard error ofthe estimate.

The above assumptions do not hold good in the theory of small samples. Thus, anew technique is needed to deal with the theory of small samples. A sample is small when it consists of less than 30 items. ($n < 30$).

Since in many of the problems it becomes necessary to take a small size sample, considerable attention has been paid in developing suitable tests for dealing with problems of small samples. The greatest contribution to the theory of small samples is that of **Sir William Gosset and Prof. R.A. Fisher**. Sir William Gosset published his discovery in 1905 under the pen name 'Student' and later on developed and extended by Prof. R.A.Fisher. He gave a test popularly known as ' t-test' .

t - statistic definition:

If $x_1, x_2, \ldots\ldots x_n$ is a random sample of size n from a normal population with mean μ and variance σ^2, then Student's t-statistic is defined as $t = \dfrac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$

where $\bar{x} = \dfrac{\sum x}{n}$ is the sample mean

and $S^2 = \dfrac{1}{n-1}\sum(x-\bar{x})^2$

is an unbiased estimate of the population variance σ^2 It follows student's t-distribution with $\nu = n-1$ d.f

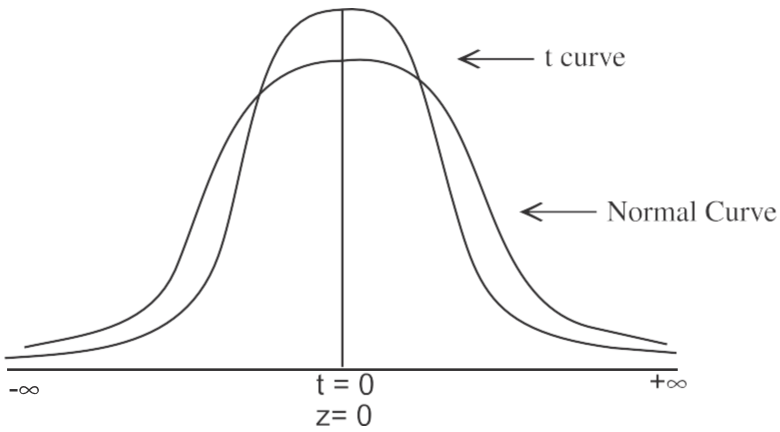
Assumptions for students t-test:

1. The parent population from which the sample drawn is **normal**.
2. The sample observations are random and **independent**.
3. The population standard deviation **σ is unknown**.

Properties of t- distribution:

1. t-distribution ranges from -∞ to +∞, just as in a normal distribution.
2. Like the normal distribution, t-distribution also symmetrical and has a mean zero.
3. t-distribution has a greater dispersion than the standard normal distribution.
4. As the sample size approaches 30, the t-distribution, approaches the Normal distribution.

Comparison between Normal curve and corresponding t - curve:



Degrees of freedom (d.f):

Suppose it is asked to write any four number then one will have all the numbers of his choice. If a restriction is applied or imposed to the choice that the sum of these number should be 50. Here, we have a choice to select any three numbers, say 10, 15, 20 and thefourth number is 5: [50 – (10 + 15 + 20)]. Thus our choice of freedom is reduced by one, on the condition that the total be 50. Therefore the restriction placed on the freedom is one and degree of freedom is three. As the restrictions increase, the freedom is reduced.

The number of independent observations which are used to calculate the statistic is known as the **degrees of freedom** and is usually denoted **by υ (Nu).**

The number of degrees of freedom for n observations is n - k where k is the number of independent linear constraint imposed upon them.

For the student' s t-distribution the number of degrees of freedom is the sample size minus one. It is denoted by υ = n -1 The degrees of freedom plays a very important role in test of a hypothesis.

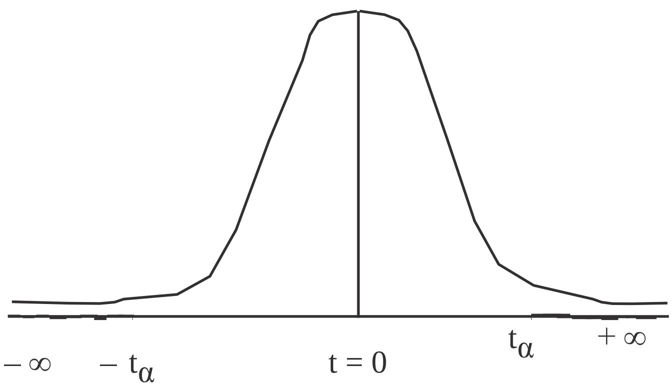
When we fit a distribution the number of degrees of freedom is (n– k-1) where n is number of observations and k is number of parameters estimated from the data.

Critical value of t:

The column figures in the main body of the table come under the headings $t_{0.10}$, $t_{0.50}$, $t_{0.025}$, $t_{0.010}$ and $t_{0.005}$. The subscripts give the proportion of the distribution in ‘ tail’ area. Thusfor two-tailed test at 5% level of significance there will be two rejection areas each containing2.5% of the total area and the required column is headed $t_{0.025}$

For one tailed test, at 5% level, the rejection area lies in one end of the tail of thedistribution and the required column is headed $t_{0.05}$.

Critical value of t – distribution



Applications of t-distribution:

The t-distribution has a number of applications in statistics,

(i) t-test for significance of single mean,

(ii) t-test for significance of the difference between two sample means,

(a) Independent samples (b) Related samples: paired t-test

[Note:- t-distribution has few more applications which are not listed here.]

4.2 Test for Single Mean

Test of Hypotheses for Normal Population Mean (Population Variance is Unknown)

Procedure:

Step 1 : Let μ and σ^2 be respectively the mean and variance of the population under study, where σ^2 is unknown. If μ_0 is an admissible value of μ, then frame the null hypothesis as

H_0: $\mu = \mu_0$ and choose the suitable alternative hypothesis from

(i) H_1: $\mu \neq \mu_0$ (ii) H_1: $\mu > \mu_0$ (iii) H_1: $\mu < \mu_0$

Step 2 : Describe the sample/data and its descriptive measures. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is small ($n < 30$).

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ under H_0, where $\bar{X}$ and S are the sample mean and sample standard deviation respectively. The approximate sampling distribution of the test statistic under H_0 is the t-distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of t for the given sample $(x_1, x_2, \ldots x_n)$ as $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$.

here $\bar{x}$ is the sample mean and $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation.

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Critical Value (t_e)	$t_{n-1,\alpha/2}$	$t_{n-1,\alpha}$	$-t_{n-1,\alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Rejection Rule	$\|t_0\| \geq t_{n-1,\alpha/2}$	$t_0 > t_{n-1,\alpha}$	$t_0 < -t_{n-1,\alpha}$

Example:

The average monthly sales, based on past experience of a particular brand of tooth paste in departmental stores is ₹ 200. An advertisement campaign was made by the company and then a sample of 26 departmental stores was taken at random and found that the average sales of the particular brand of tooth paste is ₹ 216 with a standard deviation of ₹ 8. Does the campaign have helped in promoting the sales of a particular brand of tooth paste?

Solution:

Step 1 : Hypotheses

Null Hypothesis H_0: $\mu = 200$

i.e., the average monthly sales of a particular brand of tooth paste is not significantly different from ₹ 200.

Alternative Hypothesis H_1: $\mu > 200$

i.e., the average monthly sales of a particular brand of tooth paste are significantly different from ₹ 200. It is one-sided (right) alternative hypothesis.

Step 2 : Data

The given sample information are:

Size of the sample (n) = 26. Hence, it is a small sample.

Sample mean ($\bar{x}$) = 216, Standard deviation of the sample = 8.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Since n is small, the sampling distribution of T is the t-distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of T for the given sample information is calculated from

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ as}$$

$$t_0 = \frac{216 - 200}{8/\sqrt{26}} = 10.20$$

Step 6 : Critical value

Since H_1 is one-sided (right) alternative hypothesis, the critical value at α =0.05 is

$$t_e = t_{n-1,\alpha} = t_{25,0.05} = 1.708$$

Step 7 : Decision

Since it is right-tailed test, elements of critical region are defined by the rejection rule $t_0 > t_e = t_{n-1,\alpha} = t_{25,0.05} = 1.708$. For the given sample information $t_0 = 10.20 > t_e = 1.708$. It indicates that given sample contains sufficient evidence to reject H_0. Hence, the campaign has helped in promoting the increase in sales of a particular brand of tooth paste.

Example:

A sample of 10 students from a school was selected. Their scores in a particular subject are 72, 82, 96, 85, 84, 75, 76, 93, 94 and 93. Can we support the claim that the class average scores is 90?

Solution:

Step 1 : Hypotheses

Null Hypothesis H_0: $\mu = 90$

i.e., the class average scores is not significantly different from 90.

Alternative Hypothesis H_1 : $\mu \neq 90$

i.e., the class means scores is significantly different from 90.

It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Size of the sample (n) = 10. Hence, it is a small sample.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Since n is small, the sampling distribution of T is the t - distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of T for the given sample information is calculated from $t_0 = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ as under:

x_i	$u_i = x_i - A$; $(A = 85)$	u_i^2
72	−13	169
82	−3	9
96	11	121
85	0	0
84	−1	1
75	−10	100
76	−9	81
93	8	64
94	9	81
93	8	64
	$\sum_{i=1}^{10} u_i = 0$	$\sum_{i=1}^{10} u_i^2 = 690$

Sample mean

$$\bar{x} = A + \frac{\sum_{i=1}^{10} u_i}{n} \text{ where } A \text{ is assumed mean}$$

$= 85 + 0 = 85$

Sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{10} u_i^2}$$

$$= \sqrt{\frac{1}{9} \times 690}$$

$$= \sqrt{76.67}$$

$= 8.756$

Hence,

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{85 - 90}{8.756/\sqrt{10}} = \frac{-5}{2.77}$$

$= -1.806$ and

$|t_0| = 1.806$

Step 6 : **Critical value**

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $t_e = t_{n-1, \frac{\alpha}{2}} = t_{9,0.025} = 2.262$

Step 7 : **Decision**

Since it is two-tailed test, elements of critical region are defined by the rejection rule $|t_0| > t_e = t_{n-1, \frac{\alpha}{2}} = t_{9,0.025} = 2.262$. For the given sample information $|t_0| = 1.806 < t_e = 2.262$. It indicates that given sample does not provide sufficient evidence to reject H_0. Hence, we conclude that the class average scores is 90.

4.3 Test for Equality of Means (Independent Samples)

Test of Hypotheses for Equality of Means of Two Normal Populations (Independent Random Samples)

Procedure:

Step 1 : Let μ_X and μ_Y be respectively the means of population-1 and population-2 under study. The variances of the population-1 and population-2 are assumed to be equal and unknown given by σ^2.

Frame the null hypothesis as $H_0 : \mu_X = \mu_Y$ and choose the suitable alternative hypothesis from (i) $H_1 : \mu_X \neq \mu_Y$ (ii) $H_1 : \mu_X > \mu_Y$ (iii) $H_1 : \mu_X < \mu_Y$

Step 2 : Describe the sample/data. Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2, where m and n are small (*i.e.*, $m < 30$ and $n < 30$). Here, these two samples are assumed to be independent.

Step 3 : Set up level of significance (α)

Step 4 : Consider the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ under } H_0 \ (i.e., \mu_X = \mu_Y)$$

where S_p is the "pooled" standard deviation (combined standard deviation) given by

$$S_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}};$$

and

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The approximate sampling distribution of the test statistic

$$T = \frac{(\bar{X} - \bar{Y})}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{under } H_0$$

is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$.

Step 5 : Calculate the value of T for the given sample $(x_1, x_2, \ldots x_m)$ and $(y_1, y_2, \ldots y_n)$ as

$$t_0 = \frac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Here $\bar{x}$ and $\bar{y}$ are the values of $\bar{X}$ and $\bar{Y}$ for the samples. Also $s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ are the sample variances and $s_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}}$.

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (t_e)	$t_{n-1, \frac{\alpha}{2}}$	$t_{n-1, \frac{\alpha}{2}}$	$-t_{(n-1), \alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1, \frac{\alpha}{2}}$	$t_0 > t_{n-1,\alpha}$	$t_0 < -t_{n-1,\alpha}$

Example:

The following table gives the scores (out of 15) of two batches of students in an examination.

Batch I	6	7	9	2	13	3	4	8	7	11
Batch II	5	6	5	7	1	7	2	7		

Test at 1% level of significance the average performance of the students in Batch I and Batch II are equal.

Solution:

Step 1 : Hypotheses: Let μ_X and μ_Y denote respectively the average performance of students in Batch I and Batch II. Then the null and alternative hypotheses are :

Null Hypothesis $H_0 : \mu_X = \mu_Y$

i.e., the average performance of the students in Batch I and Batch II are equal.

Alternative Hypothesis $H_1 : \mu_X \neq \mu_Y$

i.e., the average performance of the students in Batch I and Batch II are not equal.

Step 2 : Data

The given sample information are:

Sample size for Batch I : $m = 10$

Sample size for Batch II : $n = 8$

Step 3 : Level of significance

$\alpha = 1\%$

Step 4 : Test statistic

The test statistic under H_0 is

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

The sampling distribution of T under H_0 is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$

Step 5 : Calculation of test statistic

To find sample mean and sample standard deviation:

x_i	$u_i = x_i - \bar{x}$ $(\bar{x} = 7)$	u_i^2	y_i	$v_i = y_i - \bar{y}$ $(\bar{y} = 5)$	v_i^2
6	-1	1	5	0	0
7	0	0	6	1	1
9	2	4	5	0	0
2	-5	25	7	2	4
13	6	36	1	-4	16
3	-4	16	7	2	4
4	-3	9	2	-3	9
8	1	1	7	2	4
7	0	0			
11	4	16			
$\sum_{i=1}^{10} x_i = 70$	$\sum_{i=1}^{10} u_i = 0$	$\sum_{i=1}^{10} u_i^2 = 108$	$\sum_{i=1}^{8} y_i = 40$	$\sum_{i=1}^{8} v_i = 0$	$\sum_{i=1}^{8} v_i^2 = 38$

To find sample means:

Let $(x_1, x_2, ..., x_{10})$ and $(y_1, y_2, ..., y_8)$ denote the scores of students in Batch I and Batch II respectively.

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{70}{10} = 7$$

$$\bar{y} = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{40}{8} = 5$$

To find combined sample standard deviation:

$$s_X^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{10} u_i^2 = \frac{108}{9} = 12$$

$$s_Y^2 = \frac{1}{7}\sum_{i=1}^{8}(y_i - \bar{y})^2 = \frac{1}{7}\sum_{i=1}^{8} v_i^2 = \frac{38}{7} = 5.4$$

Pooled standard deviation is:

$$S_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} = \sqrt{\frac{108+38}{10+8-2}} = \sqrt{9.125} = 3.021$$

The value of T is calculated for the given information as

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{7-5}{3.021\sqrt{\frac{1}{10} + \frac{1}{8}}} = 1.3957$$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.01$ is

$t_e = t_{m+n-2,\frac{\alpha}{2}} = t_{16,0.005} = 2.921$

Step 7 : Decision

Since it is two-tailed test, elements of critical region are defined by the rejection rule $|t_0| < t_e = t_{m+n-2,\frac{\alpha}{2}} = t_{16,0.005} = 2.921$. For the given sample information $|t_0| = 1.3957 < t_e = 2.921$. It indicates that given sample contains insufficient evidence to reject H_0. Hence, the mean performance of the students in these batches are equal.

Example:

Two types of batteries are tested for their length of life (in hours). The following data is the summary descriptive statistics.

Type	Number of batteries	Average life (in hours)	Sample standard deviation
A	14	94	16
B	13	86	20

Is there any significant difference between the average life of the two batteries at 5% level of significance?

Solution:

Step 1 : Hypotheses

Null Hypothesis $H_0 : \mu_X = \mu_Y$

i.e., there is no significant difference in average life of two types of batteries *A* and *B*.

Alternative Hypothesis $H_0 : \mu_X \neq \mu_Y$

i.e., there is significant difference in average life of two types of batteries *A* and *B*. It is a two-sided alternative hypothesis

Step 2 : Data

The given sample information are :

m = number of batteries under type *A* = 14

n = number of batteries under type *B* = 13

$\bar{x}$ = Average life (in hours) of type *A* battery = 94

$\bar{y}$ = Average life (in hours) of type *B* battery = 86

s_X = standard deviation of type *A* battery =16

s_Y = standard deviation of type *B* battery = 20

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}.$$

The sampling distribution of T under H_0 is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$

Step 5 : Calculation of test statistic

Under null hypotheses H_0:

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

where s is the pooled standard deviation given by,

$$s_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}}$$

$$= \sqrt{\frac{(14-1)(16)^2 + (13-1)(20)^2}{14+13-2}} = \sqrt{325.12} = 18.03$$

The value of T is calculated for the given information as

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{94-86}{18.03\sqrt{\frac{1}{14}+\frac{1}{13}}} = \frac{8}{6.944} = 1.15$$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $t_e = t_{m+n-2, \frac{\alpha}{2}} = t_{25, 0.025} = 2.060$.

Step 7 : Decision

Since it is a two-tailed test, elements of critical region are defined by the rejection rule $|t_0| < t_e = t_{m+n-2, \frac{\alpha}{2}} = t_{25, 0.025} = 2.060$. For the given sample information $|t_0| = 1.15 < t_e = 2.060$. It indicates that given sample contains insufficient evidence to reject H_0. Hence, there is no significant difference between the average life of the two types of batteries.

4.4 Test for Equality of Means (Dependent / Paired Samples)

To test the equality of two means – paired *t*-test

Procedure:

Step 1 : Let X and Y be two correlated random variables having the distributions respectively $N(\mu_X, \sigma_X^2)$ (Population-1) and $N(\mu_Y, \sigma_Y^2)$ (Population-2). Let $D = X - Y$, then it has normal distribution $N(\mu_D = \mu_X - \mu_Y, \sigma_D^2)$.

Frame null hypothesis as

$H_0 : \mu_D = 0$

And choose alternative hypothesis from

(i) $H_1 : \mu_D \neq 0$ (ii) $H_1 : \mu_D > 0$ (iii) $H_1 : \mu_D < 0$

Step 2 : Describe the sample/data. Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2. Here, these two samples are correlated in pairs.

Step 3 : Set up level of significance (α)

Step 4 : Consider the test statistic

$$T = \frac{\bar{D}}{\frac{S}{\sqrt{n}}} \text{ under } H_0.$$

where $\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$; $D_i = X_i - Y_i$ and $S = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}}$.

The approximate sampling distribution of the test statistic T under H_0 is t - distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of T for the given data as

$$t_0 = \frac{\bar{d}}{\frac{s}{\sqrt{n}}}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$; $d_i = x_i - y_i$ (sample mean) and

$$s = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} \quad \text{(sample standard deviation)}$$

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_D \neq 0$	$\mu_D > 0$	$\mu_D < 0$
Critical Value (t_e)	$t_{n-1, \frac{\alpha}{2}}$	$t_{n-1, \alpha}$	$-t_{n-1, \alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_D \neq 0$	$\mu_D > 0$	$\mu_D < 0$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1, \frac{\alpha}{2}}$	$t_0 > t_{n-1, \alpha}$	$t_0 < -t_{n-1, \alpha}$

A company gave an intensive training to its salesmen to increase the sales. A random sample of 10 salesmen was selected and the value (in lakhs of Rupees) of their sales per month, made before and after the training is recorded in the following table. Test whether there is any increase in mean sales at 5% level of significance.

Salesman	1	2	3	4	5	6	7	8	9	10
Before	15	22	6	17	12	20	18	14	10	16
After	17	23	16	20	14	21	18	20	10	11

Solution:

Step 1 : **Hypotheses**

Null Hypothesis $H_0 : \mu_D = 0$

i.e., there is no significant increase in the mean sales after the training.

Alternative Hypothesis $H_1 : \mu_D > 0$

i.e., there is significant increase in the mean sales after the training. It is a one-sided alternative hypothesis.

Step 2 : **Data**

Sample size $n = 10$

Step 3 : **Level of significance**

$\alpha = 5\%$

Step 4 : **Test statistic**

Test statistic under the null hypothesis is

$$T = \frac{\bar{D}}{\frac{S}{\sqrt{n}}}$$

The sampling distribution of T under H_0 is t - distribution with (10–1) = 9 degrees of freedom.

Step 5 : Calculation of test statistic

To find $\bar{d}$ and s:

Let x denote sales before training and y denote sales after training

Salesmen	x_i	y_i	$d_i = y_i - x_i$	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
1	15	17	2	0	0
2	22	23	1	-1	1
3	6	16	10	8	64
4	17	20	3	1	1
5	12	14	2	0	0
6	20	21	1	-1	1
7	18	18	0	-2	4
8	14	20	6	4	16
9	10	10	0	-2	4
10	16	11	-5	-7	49
		Total	$\sum_{i=1}^{n} d_i = 20$	$\sum_{i=1}^{n}(d_i - \bar{d}) = 0$	$\sum_{i=1}^{n}(d_i - \bar{d})^2 = 140$

Here instead of $d_i = x_i - y_i$ it is assumed $d_i = y_i - x_i$ for calculations to be simpler.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{20}{10} = 2$$

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{140}{9}} = \sqrt{15.56} = 3.94$$

The calculated value of the statistic is

$$t_0 = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} = \frac{2}{\frac{3.94}{\sqrt{10}}} = 1.6052$$

Step 6 : Critical value

Since H_1 is a one-sided alternative hypothesis, the critical value at 5% level of significance is $t_e = t_{n-1,\alpha} = t_{9,0.05} = 1.833$

Step 7 : Decision

It is a one-tailed test. Since $|t_0| = 1.6052 < t_e = t_{n-1,\alpha} = t_{9,0.05} = 1.833$, H_0 is not rejected. Hence, there is no evidence that the mean sales has increased after the training.

4.5 Chi-Square Test

Karl Pearson (1857-1936) was a English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. He was the first to examine whether the observed data support a given specification, in a paper published in 1900. He called it 'Chi-square goodness of fit' test which motivated research in statistical inference and led to the development of statistics as separate discipline.


Karl Pearson

Karl Pearson chi-square test the dawn of Statistical Inference - C R Rao.

Karl Pearson's famous chi square paper appeared in the spring of 1900, an auspicious beginning to a wonderful century for the field of statistics - B. Efron

Chi-square distribution

The square of standard normal variable is known as a chi-square variable with 1 degree of freedom (d.f.). Thus

If $X \sim N(\mu, \sigma^2)$, then it is known that $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Further Z^2 is said to follow χ^2 – distribution with 1 degree of freedom (χ^2 – pronounced as chi-square)

Note: i) If $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ are n *iid* random variables, then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (X_i - \mu)/\sigma^2 \text{ follows } \chi^2 \text{ with } n \text{ d.f (additive property of } \chi^2)$$

ii) If μ is replaced by $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$ follows χ^2_{n-1}

Properties of Chi-square distribution

- It is a continuous distribution.
- The distribution has only one parameter *i.e.* n *d.f.*
- The shape of the distribution depends upon the *d.f*, n.
- The mean of the chi-square distribution is n and variance $2n$
- If U and V are independent random variables having χ^2 distributions with degree of freedom n_1 and n_2 respectively, then their sum $U + V$ has the same χ^2 distribution with *d.f* $n_1 + n_2$.

Applications / uses of Chi-square distribution

- To test the variance of the normal population, using the statistic in note (ii) (sec. 2.2.1)
- To test the independence of attributes. (sec. 2.2.5)
- To test the goodness of fit of a distribution. (sec. 2.2.6)
- The sampling distributions of the test statistics used in the last two applications are approximately chi-square distributions.

Chi-Square Test for Independence of Attributes

Attributes: Attributes are qualitative characteristic such as levels of literacy, employment status, *etc.*, which are quantified in terms of levels/scores.

Contigency table: Independence of two attributes is an important statistical application in which the data pertaining to the attributes are cross classified in the form of a two – dimensional table. The levels of one attribute are arranged in rows and of the other in columns. Such an arrangement in the form of a table is called as a contingency table.

Computational steps for testing the independence of attributes:

Step 1 : Framing the hypotheses

Null hypothesis H_0: The two attributes are independent

Alternative hypothesis H_1: The two attributes are not independent.

Step 2 : Data

The data set is given in the form of a contigency as under. Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N}; \; i = 1,2,\ldots m; \; j = 1,2,..n$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

The contingency table consisting of m rows and n columns.
The observed data is presented in the form of a contingency table :

		Attribute B						Total
		B_1	B_2	…	B_j	…	B_n	
Attribute A	A_1	O_{11}	O_{12}	…	O_{1j}	…	O_{1n}	R_1
	A_2	O_{21}	O_{22}	…	O_{2j}	…	O_{2n}	R_2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	A_i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{in}	R_i
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	A_m	O_{m1}	O_{m2}	…	O_{mj}	…	O_{mn}	R_m
Total	C_1	C_2	…	C_j	…	C_n	$N = m\times n$	

Step 3 : **Level of significance**

Fix the desired level of significance α

Step 4 : **Calculation**

Calculate the value of the test statistic as

$$\chi_0^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Step 5 : **Critical value**

The critical value is obtained from the table of χ^2 with $(m-1)(n-1)$ degrees of freedom at given level of significance, α as $\chi^2_{(m-1)(n-1),\,\alpha}$.

Step 6 : **Decision**

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value. If $\chi_0^2 \geq \chi^2_{(m-1)(n-1),\,\alpha}$ reject H_0.

Note:

- N, the total frequency should be reasonably large, say greater than 50.
- No theoretical cell-frequency should be less than 5. If cell frequencies are less than 5, then it should be grouped such that the total frequency is made greater than 5 with the preceding or succeeding cell.

Example:

The following table gives the performance of 500 students classified according to age in a computer test. Test whether the attributes age and performance are independent at 5% of significance.

Performance	Below 20	21-30	Above 30	Total
Average	138	83	64	285
Good	64	67	84	215
Total	202	150	148	500

Solution:

Step 1 : **Null hypothesis** H_0: The attributes age and performance are independent.

Alternative hypothesis H_1: The attributes age and performance are not independent.

Step 2 : **Data**

Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N} \quad i = 1,2;\ j = 1,2,3$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

Performance	Below average	Average	Above average	Total
Average	$\frac{285\times 202}{500} = 115.14$	$\frac{285\times 150}{500} = 85.5$	$\frac{285\times 148}{500} = 84.36$	285
Good	$\frac{215\times 202}{500} = 86.86$	$\frac{215\times 150}{500} = 64.5$	$\frac{215\times 148}{500} = 63.64$	215
Total	202	150	148	500

Step 3 : **Level of significance** α = 5%

Step 4 : **Calculation**

Calculate the value of the test statistic as

$$\chi_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

This chi-square test statistic is calculated as follows:

$$\chi_0^2 = \frac{(138-115.14)^2}{115.14} + \frac{(83-85.50)^2}{88.50} + \frac{(64-84.36)^2}{84.36} + \frac{(64-86.86)^2}{86.86} + \frac{(67-64.50)^2}{64.50} + \frac{(84-63.64)^2}{63.64}$$

= 22.152 with degrees of freedom (3−1)(2−1) = 2

Step 5 : **Critical value**

From the chi-square table the critical value at 5% level of significance is $\chi^2_{(2-1)(3-1),0.05} = \chi^2_{2,0.05} = 5.991$.

Step 6 : Decision

As the calculated value $\chi_0^2 = 22.152$ is greater than the critical value $\chi^2_{2,0.05} = 5.991$, the null hypothesis H_0 is rejected. Hence, the performance and age of students are not independent.

NOTE

If the contigency table is 2 x 2 then the value of χ^2 can be calculated as given below:

	A	not A	Total
B	a	b	$a+b$
not B	c	d	$c+d$
Total	$a+c$	$b+d$	$N=a+b+c+d$

$$\chi_0^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_\alpha^2(1 d.f)$$

A survey was conducted with 500 female students of which 60% were intelligent, 40% had uneducated fathers, while 30 % of the not intelligent female students had educated fathers. Test the hypothesis that the education of fathers and intelligence of female students are independent.

Solution:

Step 1 : Null hypothesis H_0: The attributes are independent *i.e.* No association between education fathers and intelligence of female students

Alternative hypothesis H_1: The attributes are not independent *i.e* there is association between education of fathers and intelligence of female students

Step 2 : Data

The observed frequencies (O) has been computed from the given information as under.

	Intelligent females	Not intelligent females	Row total
Educated fathers	$300-120 = 180$	$\frac{30}{100}\times 200 = 60$	240
Uneducated fathers	$\frac{40}{100}\times 300 = 120$	$200-60 = 140$	260
Total	$\frac{60}{100}\times 500 = 300$	$500-300 = 200$	N= 500

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where, $a = 180$, $b = 60$, $c = 120$, $d = 140$, $N = 500$

$$\chi_0^2 = \frac{500(180\times140-60\times120)^2}{(180+60)(120+140)(180+120)(60+140)} = 43.269$$

Step 5 : Critical value

From chi-square table the critical value at 5% level of significance is $\chi^2_{1,0.05} = 3.841$

Step 6 : Decision

The calculated value $\chi_0^2 = 43.269$ is greater than the critical value $\chi^2_{1,0.05} = 3.841$, the null hypothesis H_0 is rejected. Hence, education of fathers and intelligence of female students are not independent.

Chi-Square Test for Goodness of Fit

Another important application of chi-square distribution is testing goodness of a pattern or distribution fitted to given data. This application was regarded as one of the most important inventions in mathematical sciences during 20th century. Goodness of fit indicates the closeness of observed frequency with that of the expected frequency. If the curves of these two distributions do not coincide or appear to diverge much, it is noted that the fit is poor. If two curves do not diverge much, the fit is fair.

Computational steps for testing the significance of goodness of fit:

Step 1 : Framing of hypothesis

Null hypothesis H_0: The goodness of fit is appropriate for the given data set

Alternative hypothesis H_1 : The goodness of fit is not appropriate for the given data set

Step 2 : Data

Calculate the expected frequencies (E_i) using appropriate theoretical distribution such as Binomial or Poisson.

Step 3 : Select the desired level of significance α

Step 4 : Test statistic

The test statistic is

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

where $k =$ number of classes

O_i and E_i are respectively the observed and expected frequency of i^{th} class such that $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$.

If any of E_i is found less than 5, the corresponding class frequency may be pooled with preceding or succeeding classes such that E_i's of all classes are greater than or equal to 5. It may be noted that the value of k may be determined after pooling the classes.

The approximate sampling distribution of the test statistic under H_0 is the chi-square distribution with k-1-s $d.f$, s being the number of parametres to be estimated.

Step 5 : Calculation

Calculate the value of chi-square as

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The above steps in calculating the chi-square can be summarized in the form of the table as follows:

Step 6 : Critical value

The critical value is obtained from the table of χ^2 for a given level of significance α.

Step 7 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value, at the desired level of significance.

Example:

Five coins are tossed 640 times and the following results were obtained.

Number of heads	0	1	2	3	4	5
Frequency	19	99	197	198	105	22

Fit binomial distribution to the above data.

Solution:

Step 1 : Null hypothesis H_0: Fitting of binomial distribution is appropriate for the given data.

Alternative hypothesis H_1: Fitting of binomial distribution is not appropriate to the given data.

Step 2 : Data

Compute the expected frequencies:

n = number of coins tossed at a time = 5

Let X denote the number of heads (success) in n tosses

N = number of times experiment is repeated = 640

To find mean of the distribution

x	f	fx
0	19	0
1	99	99
2	197	394
3	198	594
4	105	420
5	22	110
Total	640	1617

Mean : $\bar{x} = \frac{\sum fx}{\sum f} = \frac{1617}{640} = 2.526$

The probability mass function of binomial distribution is :

$$p(x) = {}^nC_x\, p^x q^{n-x},\ x = 0,1,\ldots,n \qquad (2.1)$$

Mean of the binomial distribution is $\bar{x} = np$.

Hence, $\hat{p} = \frac{\bar{x}}{n} = \frac{2.526}{5} \approx 0.5$

$\hat{q} = 1 - \hat{p} \approx 0.5$

For $x = 0$, the equation (2.1) becomes

$P(X = 0) = P(0) = 5c_0\,(0.5)^5 = 0.03125$

The expected frequency at $x = N\,P(x)$

The expected frequency at $x = 0 : N \times P(0)$

$= 640 \times 0.03125 = 20$

We use recurrence formula to find the other expected frequencies.

The expected frequency at $x+1$ is

$\frac{n-x}{x+1}\left(\frac{p}{q}\right) \times$ Expected frequency at x

x	$\frac{n-x}{x+1}$	$\frac{p}{q}$	$\frac{n-x}{x+1}\left(\frac{p}{q}\right)$	Expected frequency at $x = N\,P(x)$
0	5	1	5	20
1	2	1	2	100
2	1	1	1	200
3	0.5	1	0.5	200
4	0.2	1	0.2	100
5	0	1	0	20

Table of expected frequencies:

Number of heads	0	1	2	3	4	5	Total
Expected frequencies	20	100	200	200	100	20	640

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

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

Step 5 : Calculation

The test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
19	20	–1	1	0.050
99	100	–1	1	0.010
197	200	–3	9	0.045
198	200	–2	4	0.020
105	100	5	25	0.250
22	20	2	4	0.200
			Total	0.575

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 0.575$$

Step 6 : Critical value

Degrees of freedom $= k - 1 - s = 6 - 1 - 1 = 4$

Critical value for *d.f* 4 at 5% level of significance is 9.488 *i.e.*, $\chi^2_{4,0.05} = 9.488$

Step 7 : Decision

As the calculated χ_0^2 (=0.575) is less than the critical value $\chi^2_{4,0.05} = 9.488$, we do not reject the null hypothesis. Hence, the fitting of binomial distribution is appropriate.

Example:

A packet consists of 100 ball pens. The distribution of the number of defective ball pens in each packet is given below:

x	0	1	2	3	4	5
f	61	14	10	7	5	3

Examine whether Poisson distribution is appropriate for the above data at 5% level of significance.

Solution:

Step 1 : **Null hypothesis** H_0: Fitting of Poisson distribution is appropriate for the given data.

Alternative hypothesis H_1: Fitting of Poisson distribution is not appropriate for the given data.

Step 2 : **Data**

The expected frequencies are computed as under:

To find the mean of the distribution.

x	f	fx
0	61	0
1	14	14
2	10	20
3	7	21
4	5	20
5	3	15
Total	100	90

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{90}{100} = 0.9$$

Probability mass function of Poisson distribution is:

$$p(x) = \frac{e^{-m} m^x}{x!};\ x = 0, 1, \ldots \qquad (2.2)$$

In the case of Poisson distribution mean $(m) = \bar{x} = 0.9$.

At $x = 0$, equation (2.2) becomes

$$p(0) = \frac{e^{-m}\ m^0}{0!} = e^{m} = e^{0.9} = 0.4066.$$

The expected frequency at x is $N\,P(x)$

Therefore, The expected frequency at 0 is

$$N \times P(0)$$
$$= 100 \times 0.4066$$
$$= 40.66$$

We use recurrence formula to find the other expected frequencies.

The expected frequency at $x+1$ is

$$\frac{m}{x+1} \times \text{Expected frequency at } x$$

x	$\frac{m}{x+1}$	Expected frequency at $x = N\,P(x)$
0	0.9	40.66
1	$\frac{0.9}{2}$	36.594
2	$\frac{0.9}{3}$	16.4673
3	$\frac{0.9}{4}$	4.94019
4	$\frac{0.9}{5}$	1.1115
5	$\frac{0.9}{6}$	0.20007

Table of expected frequency distribution (on rounding to the nearest integer)

x	0	1	2	3	4	5
Expected frequency	41	37	16	5	1	0

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

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

Step 5 : Calculation

Test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
61	41	20	400	9.756
14	37	-23	529	14.297
10	16	-6	36	2.250
7	5			
5 } 15	1 } 6	9	81	13.5
3	0			
			Total	39.803

Note: In the above table, we find the cell frequencies 0,1 in the expected frequency column (E) is less than 5, Hence, we combine (pool) with either succeeding or preceding one such that the total is made greater than 5. Here we have pooled with preceding frequency 5 such that the total frequency is made greater than 5. Correspondingly, cell frequencies in observed frequencies are pooled.

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$=39.803$

Step 6 : Critical value

Degrees of freedom = $(k - 1 - s) = 4 - 1 - 1 = 2$

Critical value for 2 $d.f$ at 5% level of significance is 5.991 *i.e.*, $\chi^2_{2,0.05} = 5.991$

Step 7 : Decision

The calculated χ_0^2(=39.803) is greater than the critical value (5.991) at 5% level of significance. Hence, we reject H_0. i.e., fitting of Poisson distribution is not appropriate for the given data.

Example:

A sample 800 students appeared for a competitive examination. It was found that 320 students have failed, 270 have secured a third grade, 190 have secured a second grade and the remaining students qualified in first grade. The general opinion that the above grades are in the ratio 4:3:2:1 respectively. Test the hypothesis the general opinion about the grades is appropriate at 5% level of significance.

Step 1 : Null hypothesis H_0: The result in four grades follows the ratio 4:3:2:1

Alternative hypothesis H_1: The result in four grades does not follows the ratio 4:3:2:1

Step 2 : Data

Compute expected frequencies:

Under the assumption on H_0, the expected frequencies of the four grades are:

$$\frac{4}{10}\times 800 = 320;\ \frac{3}{10}\times 800 = 240;\ \frac{2}{10}\times 800 = 160;\ \frac{1}{10}\times 800 = 80$$

Step 3 : Test statistic

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
320	320	0	0	0
270	240	30	900	3.75
190	160	30	900	5.625
20	80	-60	3600	45
			Total	54.375

The test statistic is calculated as

$$\chi_0^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}$$
$$= 54.375$$

Step 4 : Critical value

The critical value of χ^2 for 3 d.f. at 5% level of significance is 7.81 *i.e.*, $\chi^2_{3,0.05} = 7.81$

Step 5 : Decision

As the calculated value of χ_0^2 (=54.375) is greater than the critical value $\chi^2_{3,0.05} = 7.81$, reject H_0. Hence, the results of the four grades do not follow the ratio 4:3:2:1.

Example:

The following table shows the distribution of digits in numbers chosen at random from a telephone directory.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the occurence of the digits in the directory are equal at 5% level of significance.

Step 1 : Null hypothesis H_0: The occurrence of the digits are equal in the directory.

Alternative hypothesis H_1: The occurrence of the digits are not equal in the directory.

Step 2 : Data

The expected frequency for each digit = $\frac{10000}{10} = 1000$

Step 3 : Level of significance $\alpha = 5\%$

Step 3 : Test statistic

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1026	1000	26	676	0.676
1107	1000	107	11449	11.449
997	1000	3	9	0.009
966	1000	34	1156	1.156
1075	1000	75	5625	5.625
933	1000	67	4489	4.489
1107	1000	107	11449	11.449
972	1000	28	784	0.784
964	1000	36	1296	1.296
853	1000	147	21609	21.609
			Total	58.542

The test statistic is calculated as

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 58.542$$

Step 4 : Critical value

Critical value for 9 df at 5% level of significance is 16.919 i.e., $\chi^2_{9,0.05} = 16.919$

Step 5 : Decision

Since the calculated χ_0^2 (58.542) is greater than the critical value $\chi^2_{9,0.05} = 16.919$, reject H_0. Hence, the digits are not uniformly distributed in the directory.

4.6 F- test for Equality of Two Variances

F-Distribution and its Applications

F-statistic is the ratio of two sums of the squares of deviations of observations from respective means. The sampling distribution of the statistic is *F*-distribution.

Definition: *F*-Distribution

Let X and Y be two independent χ^2 random variates with m and n degrees of freedom respectively. Then $F = \frac{X/m}{Y/n}$ is said to follow *F*-distribution with (m, n) degrees of freedom. This *F*-distribution is named after the famous statistician R.A. Fisher (1890 to 1962).

Definition: *F*-Statistic

Let $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ be two independent random samples drawn from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ populations respectively.

Then,

$$\frac{1}{\sigma_X^2}\sum_{i=1}^{m}\left(X_i-\bar{X}\right)^2 \sim \chi^2_{m-1} \text{ and } \frac{1}{\sigma_Y^2}\sum_{j=1}^{n}\left(Y_j-\bar{Y}\right)^2 \sim \chi^2_{n-1}$$

are independent

(1) Hence, *F*-Statistic is defined as

$$F=\frac{(m-1)S_X^2}{\sigma_X^2}\Big/\frac{(n-1)S_Y^2}{\sigma_Y^2} \sim F_{m-1,n-1}$$

where

$$S_X^2=\frac{1}{m-1}\sum_{i=1}^{m}\left(X_i-\bar{X}\right)^2 \text{ and } S_Y^2=\frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j-\bar{Y}\right)^2$$

CARE

If the populations are not normal, F – test may not be used.

Assumptions for testing the ratio of two normal population variances

i) The population from which the samples were obtained must be normally distributed.

ii) The two samples must be independent of each other.

(2) *F*-Statistic is also defined as the ratio of two mean square errors.

Applications of *F*-distribution

The following are some of the important applications where the sampling distribution of the respective statistic under H_0 is *F*-distribution.

(i) Testing the equality of variances of two normal populations. [Using (1)]

(ii) Testing the equality of means of k (>2) normal populations. [Using (2)]

(iii) Carrying out analysis of variance for two-way classified data. [Using (2)]

F- test for Equality of Two Variances

Test procedure:

This test compares the variances of two independent normal populations, *viz.*, $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

Step 1 : **Null Hypothesis** $H_0 : \sigma_X^2 = \sigma_Y^2$

That is, there is no significant difference between the variances of the two normal populations.

The alternative hypothesis can be chosen suitably from any one of the following

(i) $H_1 : \sigma_X^2 < \sigma_Y^2$ (ii) $H_1 : \sigma_X^2 > \sigma_Y^2$ (iii) $H_1 : \sigma_X^2 \neq \sigma_Y^2$

Step 2 : **Data**

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent samples drawn from two normal populations respectively.

Step 3 : **Level of significance** α

Step 4 : **The test Statistic**

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$$ under H_0 and its sampling distribution under H_0 is $F_{(m-1,\, n-1)}$.

Step 5 : **Calculation of the Test Statistic**

The test statistic $F_0 = \frac{s_X^2}{s_Y^2}$

Step 6 : **Critical values**

H_1	$\sigma_X^2 < \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$
Critical value(s) f_e	$f_{(m-1,\, n-1), 1-\alpha}$	$f_{(m-1,\, n-1),\, \alpha}$	$f_{(m-1,\, n-1), 1-\alpha/2}$ and $f_{(m-1,\, n-1),\, \alpha/2}$

Step 7 : **Decision**

H_1	$\sigma_X^2 < \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$
Rejection Rule	$F_0 \leq f_{(m-1,\, n-1),\, 1-\alpha}$	$F_0 \geq f_{(m-1,\, n-1),\, \alpha}$	$F_0 \leq f_{(m-1,\, n-1),\, 1-\alpha/2}$ or $F_0 \geq f_{(m-1,\, n-1),\, \alpha/2}$

Note 1: Since $f_{(m-1,\, n-1),\, 1-\alpha}$ is not avilable in the given *F*-table, it is computed as the reciprocal of $f_{(n-1,\, m-1), \alpha}$.

i.e., $f_{(m-1,\, n-1),\, 1-\alpha} = \dfrac{1}{f_{(n-1,\, m-1),\, \alpha}}$

Note 2: A *F*-test is based on the ratio of variances, it is also known as Variance Ratio Test.

Example:

Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means as 160 inches square and 91 inches square respectively. Test the hypothesis that the variances of the two populations from which the samples are drawn are equal at 10% level of significance.

Solution:

Step 1 : Null Hypothesis: $H_0 : \sigma_X^2 = \sigma_Y^2$

That is there is no significant difference between the two population variances.

Alternative Hypothesis: $H_1 : \sigma_X^2 \neq \sigma_Y^2$

That is there is significant difference between the two population variances.

Step 2 : Data

$m = 9, n = 8$

$$\sum_{i=1}^{9}(x_i - \bar{x})^2 = 160 \qquad \sum_{j=1}^{8}(y_j - \bar{y})^2 = 91$$

Step 3 : Level of significance

$\alpha = 10\%$

Step 4 : Test Statistic $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$, under H_0.

Step 5 : Calculation

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 \text{ and } s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

$$s_X^2 = \frac{160}{8} = 20 \qquad s_Y^2 = \frac{91}{7} = 13$$

$$F_0 = \frac{s_X^2}{s_Y^2} = \frac{20}{13} = 1.54$$

Step 6 : Critical values

Since H_1 is a two-sided alternative hypothesis the corresponding critical values are:

$$f_{(8,7),0.05} = 3.73 \text{ and } f_{(8,7),0.95} = \frac{1}{f_{(7,8),0.05}} = \frac{1}{3.5} = 0.286$$

Step 7 : Decision

Since $f_{(8,7),0.95} = 0.286 < F_0 = 1.54 < f_{(8,7),0.05} = 3.73$, the null hypothesis is not rejected and we conclude that there is no significant difference between the two population variances.

Example:

A medical researcher claims that the variance of the heart rates (in beats per minute) of smokers is greater than the variance of heart rates of people who do not smoke. Samples from two groups are selected and the data is given below. Using = 0.05, test whether there is enough evidence to support the claim.

Smokers	Non Smokers
$m = 25$	$n = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Solution:

Step 1 : Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

That is there is no significant difference between the two population variances.

$H_1: \sigma_1^2 > \sigma_2^2$

That is, the variance of heart rates of smokers is greater than that of non-smokers.

Step 2 : Data

Smokers	Non Smokers
$m = 25$	$n = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$$

Step 5 : Calculation

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

Step 6 : Critical value

$$f_{(m-1,n-1),0.05} = f_{(24,17),0.05} = 2.19$$

Step 7 : Decision

Since $F_0 = 3.6 > f_{(24,17),0.05} = 2.19$, the null hypothesis is rejected and we conclude that the variance of heart beats for smokers seems to be considerably higher compared to that of the non-smokers.

The following table gives the random sample of marks scored by students in two schools, A and B.

School A	63	72	80	60	85	83	70	72	81
School B	86	93	64	82	81	75	86	63	63

Is the variance of the marks of students in school A is less than that of those in school B? Test at 5% level of significance.

Solution:

Let $X_1, X_2, \ldots, X_m$ represent sample values for school A and let $Y_1, Y_2, \ldots, Y_n$ represent sample values for school B.

Step 1 : **Null Hypothesis:** $H_1 : \sigma_X^2 = \sigma_Y^2$

That is, there is no significant difference between the two population variances.

Alternative Hypothesis: $H_1 : \sigma_X^2 < \sigma_Y^2$

That is, the variance of marks in school A is significantly less than that of school B.

Step 2 : **Data**

$X_1, X_2, \ldots, X_m$ are sample from school A

$Y_1, Y_2, \ldots, Y_n$ are sample from school B

Step 3 : **Test statistic**

$$F = \frac{s_X^2}{s_Y^2}$$

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

$$s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

Step 4 : **Calculations**

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
63	-11	121	86	9	81
72	-2	4	93	16	256
80	6	36	64	-13	169
60	-14	196	82	5	25
85	11	121	81	4	16
83	9	81	75	-2	4
70	-4	16	86	9	81
72	-2	4	63	-14	196
81	7	49	63	-14	196
666		628	693		1024

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m} = \frac{666}{9} = 74$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{693}{9} = 77$$

$$s_X^2 = \frac{1}{9-1} \times 628 = \frac{1}{8} \times 628 = 78.5$$

$$s_Y^2 = \frac{1}{9-1} \times 1024 = \frac{1}{8} \times 1024 = 128$$

$$F_0 = \frac{78.5}{128} = 0.613$$

Step 5 : Level of significance

$\alpha = 5\%$

Step 6 : Critical value

$$f_{(9-1,9-1),0.95} = \frac{1}{f_{(8,8),0.05}} = \frac{1}{3.44} = 0.291$$

Step 7 : Decision

Since $F_0 = 0.613 > f_{(8,8),0.95} = 0.291$, the null hypothesis is not rejected and we conclude that in school B there seems to be more variance present than in school A.