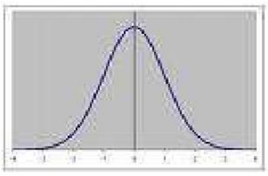
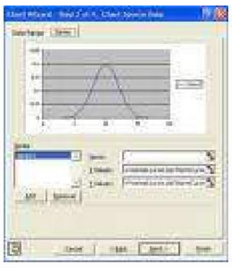
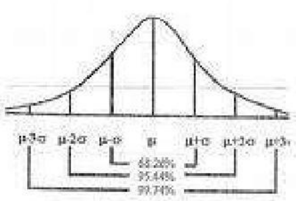
DR ATHE RAMESH

STATISTICS







Karl Pearson	**R. A. Fisher**
 Karl Pearson (né Carl Pearson)	 Sir Ronald Aylmer Fisher (1890-1962)
Born: 27 March 1857 Islington, London, England	**Born:** 17 February 1890 East Finchley, London , England

Testing of Hypothesis

Introduction: The estimate based on sample values do not equal to the true value in the population due to inherent variation in the population. The samples drawn will have different estimates compared to the true value. It has to be verified that whether the difference between the sample estimate and the population value is due to sampling fluctuation or real difference. If the difference is due to sampling fluctuation only it can be safely said that the sample belongs to the population under question and if the difference is real we have every reason to believe that sample may not belong to the

population under question. The following are a few technical terms in this context.

Hypothesis: The assumption made about any unknown characteristics is called hypothesis. It may or may be true.

Ex: 1. $\mu = 2.3$; μ be the population mean

2. $\sigma = 2.1$; σ be the population standard deviation

Population follows Normal Distribution. There are two types of hypothesis, namely null hypothesis and alternative hypothesis.

Null Hypothesis: Null hypothesis is the statement about the parameters. Such a hypothesis, which is usually a hypothesis of no difference is called null hypothesis and is usually denoted by $H_{0.}$

or

any statistical hypothesis under test is called null hypothesis. It is denoted by H_0.

1. H_0: $\mu = \mu_0$
2. H_0: $\mu_1 = \mu_2$

Alternative Hypothesis: Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis, usually denoted by H_1.

Ex:1. $H_{1:}$ μ # μ_0

2. H_1: μ_1 # μ_2

Parameter: A characteristics of population values is known as parameter. For example, population mean (μ) and population variance (σ^2).

In practice, if parameter values are not known and the estimates based on the sample values are generally used.

Statistic: A Characteristics of sample values is called a statistic. For example, sample mean ($\bar{x}$), sample variance (s^2) where $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$

$$\text{and } s^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Sampling distribution: The distribution of a statistic computed from all possible samples is known as sampling distribution of that statistic.

Standard error: The standard deviation of the sampling distribution of a statistic is known as its standard error, abbreviated as S.E.

S.E.($\bar{x}$) $= \frac{\sigma}{\sqrt{n}}$; where σ = population standard deviation and n = sample size

Sample: A finite subset of statistical objects in a population is called a sample and the number of objects in a sample is called the sample size.

Population: In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to objects belonging to a group. This group of objects under study is called population or universe.

Random sampling: If the sampling units in a population are drawn independently with equal chance, to be included in the sample then the sampling will be called random sampling. It is also referred as simple random sampling and denoted as SRS. Thus, if the population consists of „N‟ units the chance of selecting any unit is 1/N. A theoretical definition of SRS is as follows: Suppose we draw a sample of size „n‟ from a population size N; then there are $\binom{N}{n}$ possible samples of size „n‟. If all possible samples have an equal chance, $\frac{1}{\binom{N}{n}}$ of being drawn, then the sampling is said to be simple random sampling.

Simple Hypothesis: A hypothesis is said to be simple if it completely specifies the distribution of the population. For instance, in case of normal population with mean μ

and standard deviation σ, a simple null hypothesis is of the form H_0: $\mu = \mu_0$, σ is known, knowledge about μ would be enough to understand the entire distribution. For such a test, the probability of committing the type-1 error is expressed as exactly α.

Composite Hypothesis: If the hypothesis does not specify the distribution of the population completely, it is said to be a composite hypothesis. Following are some examples:

$H_0 : \mu \leq \mu_0$ and σ is known

$H_0 : \mu \geq \mu_0$ and σ is known

All these are composite because none of them specifies the distribution completely.
Hence, for such a test the LOS is specified not as α but as „at most α".

Types of Errors: In testing of statistical hypothesis there are four possible types of decisions

1. Rejecting H_0 when H_0 is true
2. Rejecting H_0 when H_0 is false
3. Accepting H_0 when H_0 is true
4. Accepting H_0 when H_0 is false

1 and 4th possibilities leads to error decisions. Statistician gives specific names to these concepts namely Type-I error and Type-II error respectively.

the above decisions can be arranged in the following table

	H_0 is true	H_0 is false
Rejecting H_0	Type-I error (Wrong decision)	Correct
Accepting H_0	Correct	Type-II error

Type-I error: Rejecting H_0 when H_0 is true

Type-II error: Accepting H_0 when H_0 is false

The probabilities of type-I and type-II errors are denoted by α and β respectively.

Degrees of freedom: It is defined as the difference between the total number of items and the total number of constraints.

If „n" is the total number of items and „k" the total number of constraints then the degrees of freedom (d.f.) is given by d.f. = n-k

Level of significance(LOS): The maximum probability at which we would be willing to risk a type-I error is known as level of significance or the size of Type-I error is level of significance. The level of significance usually employed in testing of hypothesis are 5%

and 1%. The Level of significance is always fixed in advance before collecting the sample information. LOS 5% means the results obtained will be true is 95% out of 100 cases and the results may be wrong is 5 out of 100 cases.

Critical value: while testing for the difference between the means of two populations, our concern is whether the observed difference is too large to believe that it has occurred just by chance. But then the question is how much difference should be treated as too large? Based on sampling distribution of the means, it is possible to define a cut-off or threshold value such that if the difference exceeds this value, we say that it is not an occurrence by chance and hence there is sufficient evidence to claim that the means are different. Such a value is called the critical value and it is based on the level of significance.

Steps involved in test of hypothesis:

1. The null and alternative hypothesis will be formulated
2. Test statistic will be constructed
3. Level of significance will be fixed
4. The table (critical) values will be found out from the tables for a given level of significance
5. The null hypothesis will be rejected at the given level of significance if the value of test statistic is greater than or equal to the critical value. Otherwise null hypothesis will be accepted.
6. In the case of rejection the variation in the estimates will be called „significant‟ variation. In the case of acceptance the variation in the estimates will be called „not-significant‟.

STANDARD NORMAL DEVIATE TESTS

OR

LARGE SAMPLE TESTS

If the sample size n >30 then it is considered as large sample and if the sample size n< 30 then it is considered as small sample.

SND Test or One Sample (Z-test)

Case-I: Population standard deviation (σ) is known

Assumptions:

1. Population is normally distributed
2. The sample is drawn at random

Conditions:

1. Population standard deviation σ is known
2. Size of the sample is large (say n > 30)

Procedure: Let $x_{1,}$ x_2, ...$x_{,n}$ be a random sample size of n from a normal population with mean μ and variance σ^2 .

Let $\bar{x}$ be the sample mean of sample of size „n"

Null hypothesis (H_0): population mean (μ) is equal to a specified value

μ_0 i.e. $H_0 : \mu = \mu_0$

Under H_0, the test statistic is

$$Z = \frac{\left|\bar{x} - \mu_0\right|}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

i.e the above statistic follows Normal Distribution with mean „0" and varaince"1".

If the calculated value of $|Z|$< table value of Z at 5% level of significance, H_0 is accepted and hence we conclude that there is no significant difference between the population mean and the one specified in H_0 as μ_0.

Case-II: If σ is not known

Assumptions:

1. Population is normally distributed
2. Sample is drawn from the population should be random
3. We should know the population mean

Conditions:

1. Population standard deviation σ is not known
2. Size of the sample is large (say n > 30)

Null hypothesis (H_0) : $\mu = \mu_0$

under H_0, the test statistic

$$Z = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} \sim N(0,1) \qquad \text{where } s = \sqrt{\frac{1}{n}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}$$

$\bar{x}$ = Sample mean; n = sample size

If the calculated value of Z < table value of Z at 5% level of significance, H_0 is accepted and hence we conclude that there is no significant difference between the population mean and the one specified in H_0 otherwise we do not accept H_0.

The table value of Z at 5% level of significance = 1.96 and table value of Z at 1% level of significance = 2.58.

Two sample Z-Test or Test of significant for difference of means

Case-I: when σ is known

Assumptions:

1. Populations are distributed normally
2. Samples are drawn independently and at random

Conditions:

1. Populations standard deviation σ is known
2. Size of samples are large

Procedure: Let $\bar{x_1}$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2

Let $\bar{x_2}$ be the mean of a random sample of size n_2 from another population with mean μ_2 and variance $\sigma_2{}^2$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

i.e. The null hypothesis states that the population means of the two samples are identical. Under the null hypothesis the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} \sim N(0,1) \rightarrow (1)$$

i.e the above statistic follows Normal Distribution with mean „0‟ and varaince‟1‟.

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say) i.e both samples have the same standard deviation then the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \rightarrow (2)$$

If the calculated value of $|Z|<$ table value of Z at 5% level of significance, H_0 is accepted otherwise rejected.

If H_0 is accepted means, there is no significant difference between two population means of the two samples are identical.

Example: The Average panicle length of 60 paddy plants in field No. 1 is 18.5. cms and that of 70 paddy plants in field No. 2 is 20.3 cms. With common S.D. 1.15 cms. Test whether there is significant difference between tow paddy fields w.r.t panicle length.

Solution:

Null hypothesis: H_0: There is no significant difference between the means of two paddy fields w.r.t. panicle length.

H_0: $\mu_1 = \mu_2$

Under H_0, the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{------------ (1)}$$

Where $\overline{x_1}$ = first field sample mean = 18.5 inches

$\overline{x_2}$ = second field sample mean = 20.3 inches

n_1 = first sample size = 60 n_2

= second sample size = 70

σ = common S.D. = 1.15 inches

Substitute the given values in equation (1), we get

$$Z = \frac{|18.5 - 20.3|}{1.15\sqrt{\frac{1}{60} + \frac{1}{70}}} = \frac{1.8}{1.15x0.176} = 8.89$$

Calculated value of $|Z| = 5.1$

Cal. Value of $|Z| >$ table value of Z at 5% LOS(1.96), H_0 is rejected. This means, there is highly significant difference between two paddy fields w.r.t. panicle length.

Case-II: when σ is not known

Assumptions:

1. Populations are normally distributed
2. Samples are drawn independently and at random

Conditions:

1. Population standard deviation σ is not known
2. Size of samples are large

Null hypothesis $H_0 : \mu_1 = \mu_2$

Under the null hypothesis the test statistic becomes

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

Where $\bar{x}_1$ = 1st sample mean , $\bar{x}_2$ = 2nd sample mean

s_1^2 = 1st sample variance, s_2^2 = 2nd sample variance

n_1 = 1st sample size, n_2 = 2nd sample size

If the calculated value of $|Z|$ < table value of Z at 5% level of significance, H_0 is accepted otherwise rejected.

Example:

A breeder claims that the number of filled grains per panicle is more in a new variety of paddy ACM.5 compared to that of an old variety ADT.36. To verify his claim a random sample of 50 plants of ACM.5 and 60 plants of ADT.36 were selected from the experimental fields. The following results were obtained:

(ForACM.5)	(For ADT.36)
$\bar{x}_1$ = 139.4-grains/panicle	$\bar{x}_2$ = 112.9 grains/panicle
s_1 = 26.864	s_2 = 20.1096
n_1 = 50	n_2 = 60

Test whether the claim of the breeder is correct.

Sol: Null hypothesis $H_0 : \mu_1 = \mu_2$

(i.e. the average number of filled grains per panicle is the same for both ACM.5 and ADT.36)

Under H_0, the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1) \text{ ------------ (1)}$$

Where $\overline{x_1}$ = first variety sample mean = 139.4 grains/panicle

$\overline{x_2}$ = Second variety sample mean = 112.9 grains/panicle

s_1= first sample standard deviation = 26.864

s_2= second sample standard deviation = 20.1096

n_1 = first sample size = 50

and n_2 = second sample size =

60 Substitute the given values in equation (1), we get

$$Z = \frac{|139.4 - 112.9|}{\sqrt{\frac{(26.864)^2}{50} + \frac{(20.1096)^2}{60}}}$$

$$= \frac{|26.5|}{\sqrt{14.4335 + 6.7399}}$$

$$= 4.76$$

Calculated value of Z > Table value of Z at 5% LOS (1.96), H_0 is rejected. We conclude that the number of filled grains per panicle is significantly greater in ACM.5 than in ADT.36

SMALL SAMPLE TESTS

The entire large sample theory was based on the application of "normal test". However, if the sample size „n" is small, the distribution of the various statistics, e.g., Z = $\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}}$ are far from normality and as such „normal test" cannot be applied if „n" is small.

In such cases exact sample tests, we use t-test pioneered by W.S. Gosset (1908) who wrote under the pen name of student, and later on developed and extended by Prof. R.A. Fisher.

Student's t-test: Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be a random sample of size „n" has drawn from a normal population with mean μ and variance σ^2 then student"s t – is defined by the statistic

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

$$\text{where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left\{\sum_{i=1}^{n} x_i\right\}^2}{n}\right)$$

this test statistic follows a t – distribution with (n-1) degrees of freedom (d.f.). To get the critical value of t we have to refer the table for t-distribution against (n-1) d.f. and the specific level of significance. Comparing the calculated value of t with critical value, we can accept or reject the null hypothesis.

The Range of t – distribution is - ∞ to + ∞ .

One Sample t – test

One sample t-test is a statistical procedure that is used to know the population mean and the known value of the population mean. In one sample t-test, we know the population mean. We draw a random sample from the population and then compare the sample mean with population mean and make a statistical decision as to whether or not the sample mean is different from the population mean. In one sample t-test, sample size should be less than 30.

Assumptions: 1. Population is normally distributed

2. Sample is drawn from the population and it should be random

3. We should know the population mean

Conditions: 1. Population S.D. σ is not known 2.

Size of the sample is small (<30).

Procedure: Let : Let x_1, x_2, …….,x_n be a random sample of size „n‟ has drawn from a normal population with mean μ and variance σ^2.

Null hypothesis (H_0): population mean (μ) is equal to a specified value

μ_0 i.e. H_0: $\mu = \mu_0$

Under H_0, the test statistic becomes

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

and follows student‟s t – distribution with (n-1) degrees of freedom.

$$\text{where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

We now compare the calculated value of t with the tabulated value at certain level of significance

If calculated value of |t| < table value of t at (n-1) d.f., the null hypothesis is accepted and hence we conclude that there is no significant difference between the population mean and the one specified in H_0 as μ_0 .

Example: Based on field experiments, a new variety of greengram is expected to give an yield of 13 quintals per hectare. The variety was tested on 12 randomly selected farmer fields. The yields (quintal/hectare) were recorded as 14.3, 12.6, 13.7, 10.9,13.7, 12.0, 11.4, 12.0, 13.1, 12.6, 13.4 and 13.1. Do the results conform the expectation?

Solution:

Null Hypothesis: $H_0 : \mu = \mu_0 = 13$

i.e. the results conform the expectation The test statistic becomes

$$t = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} \sim t \text{ (n-1) d.f.}$$

Where $\bar{x} = \frac{\sum x}{n}$ and $s = \sqrt{\frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}$ is an unbiased estimate of σ

Let yield $= x_i$ (say)

x_i	x_i^2
14.3	204.49
12.6	158.76
13.7	187.69
10.9	118.81
13.7	187.69
12	144
11.4	129.96
12	144
13.1	171.61
12.6	158.76
13.4	179.56
13.1	171.61
$\Sigma x = 152.8$	$\Sigma x^2 = 1956.94$

$$\bar{x} = \frac{152.8}{12} = 12.73$$

$$s = \sqrt{\frac{1}{11}\left(1956.94 - \frac{(152.8)^2}{12}\right)}$$

$$= 1.01$$

$$t = \frac{\left|12.73 - 13\right|}{\frac{1.01}{\sqrt{12}}} = \frac{0.27}{0.29} = 0.93 \text{ qa/h.}$$

t-table value at (n-1) = 11d.f. at 5 percent level of significance is 2.20.

Calculated value of t < table value of t, H_0 is accepted and we may conclude that the results conform to the expectation.

t-test for Two Samples

Assumptions: 1. Populations are distributed normally

2. Samples are drawn independently and at random

Conditions: 1. Standard deviations in the populations are same and not known

2. Size of the sample is small

Procedure: If two independent samples x_i ($i = 1,2,\ldots,n_1$) and y_j ($j = 1,2, \ldots,n_2$) of sizes $n_1$and n_2 have been drawn from two normal populations with means μ_1 and μ_2 respectively.

Null hypothesis $H_0 : \mu_1 = \mu_2$

The null hypothesis states that the population means of the two groups are identical, so their difference is zero.

Under H_0, the test statistics is $t = \dfrac{|\bar{x}-\bar{y}|}{S\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$

Where $S^2 = \dfrac{1}{n_1+n_2-2}\left[\sum x^2 - \dfrac{(\sum x)^2}{n_1} + \sum y^2 - \dfrac{(\sum y)^2}{n_2}\right]$

or

$$S^2 = \text{pooled variance} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

where s_1^2 and s_2^2 are the variances of the first and second samples respectively.

and $\bar{x} = \dfrac{\sum_{i=1}^{n_1} x_i}{n_1}$ and $\bar{y} = \dfrac{\sum_{j=1}^{n_2} y_j}{n_2}$; where $\bar{x}$ and $\bar{y}$ are the two sample means.

Which follows Student‟‟s t – distribution with (n_1+n_2-2) d.f.

If calculated value of |t| < table value of t with (n_1+n_2-2) d.f. at specified level of significance, then the null hypothesis is accepted otherwise rejected.

Example: Two verities of potato plants (A and B) yielded tubes are shown in the following table. Does the mean number of tubes of the variety „A‟‟ significantly differ from that of variety B?

Tuber yield, kg/plant

Variety-A	2.2	2.5	1.9	2.6	2.6	2.3	1.8	2.0	2.1	2.4	2.3
Variety-B	2.8	2.5	2.7	3.0	3.1	2.3	2.4	3.2	2.5	2.9	

Solution:

Hypothesis $H_0 : \mu_1 = \mu_2$

i.e the mean number of tubes of the variety „A‟‟ significantly differ from the variety „B‟‟

Statistic $t = \dfrac{|\bar{x}-\bar{y}|}{S\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \sim t\,(n_1+n_2-2)d.f$

n_1 = 1st sample size; n_2 = 2nd sample size

$\bar{x}$ = Mean of the first sample; $\bar{y}$ = mean of the second sample

$$\bar{x} = \frac{\sum x}{n_1} \quad \text{and} \quad \bar{y} = \frac{\sum y}{n_2}$$

$$\text{Where } S^2 = \frac{1}{n_1 + n_2 - 2}\left[\left\{\sum x^2 - \frac{(\sum x)^2}{n_1}\right\} + \left\{\sum y^2 - \frac{(\sum y)^2}{n_2}\right\}\right]$$

x	y	x^2	y^2
2.20	2.80	4.84	7.84
2.50	2.50	6.25	6.25
1.90	2.70	3.61	7.29
2.60	3.00	6.76	9.00
2.60	3.10	6.76	9.61
2.30	2.30	5.29	5.29
1.80	2.40	3.24	5.76
2.00	3.20	4.00	10.24
2.10	2.50	4.41	6.25
2.40	2.90	5.76	8.41
2.30	Σy = 27.40	5.29	Σy^2 = 75.94
Σx = 24.70		Σx^2 = 56.21	

$$\bar{x} = \frac{24.70}{11} \qquad \bar{y} = \frac{27.40}{10}$$

$$= 2.25 \text{ Kg} \qquad = 2.74 \text{ Kg}$$

$$\text{Where } S^2 = \frac{1}{11 + 10 - 2}\left[\left\{56.21 - \frac{(24.70)^2}{11}\right\} + \left\{75.94 - \frac{(27.40)^2}{10}\right\}\right]$$

$$= \frac{1}{19}\left[\{56.21 - 55.46\} + \{75.94 - 75.07\}\right]$$

$$= 0.09 \text{ Kg}^2$$

$$S = \sqrt{S^2} = 0.3 \text{ Kg.}$$

Test statistic $t = \dfrac{|2.25 - 2.74|}{0.3\sqrt{\left(\frac{1}{11} + \frac{1}{10}\right)}}$

$$= \frac{0.49}{0.13} = 3.77$$

Calculated value of t = 3.77

Table value of t for 19 d.f. at 5 % level of significance is 2.09

Since the calculated value of t > table value of t, the null hypothesis is rejected and hence we conclude that the mean number of tubes of the variety „A" significantly not differ from the variety „B"

Paired t – test

The paired t-test is generally used when measurements are taken from the same subject before and after some manipulation such as injection of a drug. For example, you can use a paired t test to determine the significance of a difference in blood pressure before and after administration of an experimental pressor substance.

Assumptions: 1. Populations are distributed normally

2. Samples are drawn independently and at random

Conditions: 1. Samples are related with each other

2. Sizes of the samples are small and equal

3. Standard deviations in the populations are equal and not known

Hypothesis $H_0: \mu_d = 0$

Under H_0, the test statistic becomes,

$$t = \frac{|\bar{d}|}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)} \text{ d.f.}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} a_i}{n}$ and $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} a_i^2 - \frac{\left(\sum_{i=1}^{n} a_i\right)^2}{n}\right]$

where S^2 is variance of the deviations

n = sample size; where $d_i = x_i - y_i$ (i = 1,2,……,n)

If calculated value of |t| < table value of t for (n-1)d.f. at α% level of significance, then the null hypothesis is accepted and hence we conclude that the two samples may belong to the same population. Otherwise, the null hypothesis rejected.

Example: The average number of seeds set per pod in Lucerne were determined for top flowers and bottom flowers in ten plants. The values observed were as follows:

Top flowers	4.2	5.0	5.4	4.3	4.8	3.9	4.2	3.1	4.4	5.8
Bottom flowers	4.6	3.5	4.8	3.0	4.1	4.4	3.6	3.8	3.2	2.2

Test whether there is any significant difference between the top and bottom flowers with respect to average numbers of seeds set per pod.

Solution:

Null Hypothesis $H_0: \mu_d = 0$

Under H_0 becomes, the test statistic is

$$t = \frac{|\bar{d}|}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}\, d.f.$$

Where $\bar{d} = \frac{\sum d}{n}$ and $s^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{\left(\sum d\right)^2}{n}\right]$

x	y	d=x-y	d^2
4.2	4.6	-0.40	0.16
5.0	3.5	1.50	2.25
5.4	4.8	0.60	0.36
4.3	3.0	1.30	1.69
4.8	4.1	0.70	0.49
3.9	4.4	-0.50	0.25
4.2	3.6	0.60	0.36
3.1	3.8	-0.70	0.49
4.4	3.2	1.20	1.44
5.8	2.2	3.60	12.96
		Σd = 7.90	Σd^2 = 20.45

$$\bar{d} = \frac{\sum d}{n} = \frac{7.90}{10}$$

$$= 0.79$$

$$s^2 = \frac{1}{9}\left[20.45 - \frac{(7.90)^2}{10}\right]$$

$$= 2.27$$

$$s = \sqrt{s^2} = \sqrt{2.27}$$

$$= 1.51$$

$$t = \frac{0.79}{\frac{1.51}{\sqrt{10}}}$$

$$= 1.65$$

Calculated value of t = 1.65

Table value of t for 9 d.f. at 5% level of significance is 2.26

Calculated value of t < table value of t, the null hypothesis is accepted and we conclude that there is no significant difference between the top and bottom flowers with respect to average numbers of seeds set per pod.

F – Test

In agricultural experiments the performance of a treatment is assessed not only by its mean but also by its variability. Hence, it is of interest to us to compare the variability of two populations. In testing of hypothesis the equality of variances, the greater variance is always placed in the Numerator and smaller variance is placed in the denominator.

F- test is used to test the equality of two population variances, equality of several regression coefficients, ANOVA .

F- test was discovered by G.W. Snedecor. The range of F : 0 to ∞

Let $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ be the two independent random samples of sizes n_1 and n_2 drawn from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. S_1^2 and S_2^2 are the sample variances of the two samples.

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$

Under H_0, the test statistic becomes

$$F = \frac{S_1^2}{S_2^2} \quad \text{where, } S_1^2 > S_2^2$$

Which follows F-distribution with (n_1-1, n_2-1)d.f.

Where $$S_1^2 = \frac{1}{n_1 - 1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n_1}\right] \text{ and } S_2^2 = \frac{1}{n_2 - 1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n_2}\right]$$

or $$F = \frac{S_2^2}{S_1^2} \quad \text{where } S_2^2 > S_1^2$$

Which follows F-distribution with (n_2-1, n_1-1)d.f.

If calculated value of F < table value of F with (n_2-1, n_1-1)d.f at specified level of significance, then the null hypothesis is accepted and hence we conclude that the variances of the populations are homogeneous otherwise heterogeneous.

Example: The heights in meters of red gram plants with two types of irrigation in two fields are as follows:

Tap water (x)	3.5	4.2	2.8	5.2	1.7	2.6	3.5	4.2	5.0	5.2
Saline water (y)	1.9	2.6	2.3	4.3	4.0	4.2	3.8	2.9	3.7	

Test whether the variances of the two system of irrigation are homogeneous.

Solution:

H_0: The variances of the two systems of irrigation are homogeneous.

i.e. $\sigma_1^2 = \sigma_2^2$

Under H_0, the test statistic becomes

$$F = \frac{S_1^2}{S_2^2} \quad ; (S_1^2 > S_2^2)$$

Where S_1^2 = fist sample variance = $\frac{1}{n_1 - 1}\left(\sum x^2 - \frac{(\sum x)^2}{n_1}\right)$

and S_2^2 = second sample variance = $\frac{1}{n_2 - 1}\left(\sum y^2 - \frac{(\sum y)^2}{n_2}\right)$

x	y	x^2	y^2
3.5	1.9	12.25	3.61
4.2	2.6	17.64	6.76
2.8	2.3	7.84	5.29
5.2	4.3	27.04	18.49
1.7	4	2.89	16
2.6	4.2	6.76	17.64
3.5	3.8	12.25	14.44
4.2	2.9	17.64	8.41
5	3.7	25	13.69
5.2	$\sum y = 29.7$	27.04	$\sum y^2 = 104.33$
$\sum x = 37.9$		$\sum x^2 = 156.35$	

$$S_1^2 = \frac{1}{9}\left(156.35 - \frac{(37.9)^2}{10}\right) = 1.41 \text{ mt}^2$$

and

$$S_2^2 = \frac{1}{8}\left(104.33 - \frac{(29.7)^2}{9}\right) = 0.79 \text{ mt}^2$$

$$F = \frac{S_1^2}{S_2^2} = \frac{1.41}{0.79} = 1.78$$

F calculated value = 1.78

Table value of $F_{0.05}$ for (n_1-1, n_2-1) d.f. = 3.23

Calculated value of F < Table value of at 5% level of significance, H_0 is accepted and hence we conclude that the variances of the two systems of irrigation are homogeneous.

$$\frac{1}{F_i} = F_2 \quad \text{or} \quad F_1 = \frac{1}{F_2}$$

Chi-square (χ^2) test

The various tests of significance studied earlier such that as Z-test, t-test, F-test were based on the assumption that the samples were drawn from normal population. Under this assumption the various statistics were normally distributed. Since the procedure of testing the significance requires the knowledge about the type of population or parameters of population from which random samples have been drawn, these tests are known as parametric tests.

But there are many practical situations in which the assumption of any kind about the distribution of population or its parameter is not possible to make. The alternative technique where no assumption about the distribution or about parameters of population is made are known as non-parametric tests. Chi-square test is an example of the non parametric test. Chi-square distribution is a distribution free test.

If $X_i \rightarrow N(0,1)$ then $\sum x_i^2 \sim \chi^2_n$

Chi-square distribution was first discovered by Helmert in 1876 and later independently by Karl Pearson in 1900. The range of chi-square distribution is 0 to ∞.

Measuremental data: the data obtained by actual measurement is called measuremental data. For example, height, weight, age, income, area etc.,

Enumeration data: the data obtained by enumeration or counting is called enumeration data. For example, number of blue flowers, number of intelligent boys, number of curled leaves, etc.,

χ^2 – test is used for enumeration data which generally relate to discrete variable where as t-test and standard normal deviate tests are used for measure mental data which generally relate to continuous variable.

χ^2 –test can be used to know whether the given objects are segregating in a theoretical ratio or whether the two attributes are independent in a contingency table.

The expression for χ^2 –test for goodness of fit

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

where O_i = observed frequencies

E_i = expected frequencies

n = number of cells(or classes)

Which follows a chi-square distribution with (n-1) degrees of freedom

The null hypothesis H_0 = the observed frequencies are in agreement with the expected frequencies

If the calculated value of χ^2 < Table value of χ^2 with (n-1) d.f. at specified level of significance (α), we accept H_0 otherwise we do not accept H_0.

Conditions for the validity of χ^2 –test:

The validity of χ^2-test of goodness of fit between theoretical and observed, the following conditions must be satisfied.

i) The sample observations should be independent

ii) Constraints on the cell frequencies, if any, should be linear $\Sigma o_i = \Sigma e_i$

iii) N, the total frequency should be reasonably large, say greater than 50

iv) If any theoretical (expected) cell frequency is < 5, then for the application of chi-square test it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

Applications of Chi-square Test:

1. testing the independence of attributes
2. to test the goodness of fit
3. testing of linkage in genetic problems
4. comparison of sample variance with population variance
5. testing the homogeneity of variances
6. testing the homogeneity of correlation coefficient

Test for independence of two Attributes of (2x2) Contingency Table:

A characteristic which can not be measured but can only be classified to one of the different levels of the character under consideration is called an attribute.

2x2 contingency table: When the individuals (objects) are classified into two categories with respect to each of the two attributes then the table showing frequencies distributed over 2x2 classes is called 2x2 contingency table.

Suppose the individuals are classified according to two attributes say intelligence (A) and colour (B). The distribution of frequencies over cells is shown in the following table.

A B	A1	A2	Row totals
B_1	a	B	$R_1+(a+b)$
B_2	c	D	$R_2 = (c+d)$
Column total	$C_1 = (a+c)$	$C_2 = (b+d)$	$N = (R_1+R_2)$ or (C_1+C_2)

Where R_1 and R_2 are the marginal totals of 1st row and 2nd row

C_1 and C_2 are the marginal totals of 1st column and 2nd column

N = grand total

The null hypothesis H_0: the two attributes are independent (if the colour is not dependent on intelligent)

Based on above H_0, the expected frequencies are calculated as follows.

$$E(a) = \frac{R_1 x C_1}{N} \;;\quad E(b) = \frac{R_1 x C_2}{N};\; E(c) = \frac{R_2 x C_1}{N} \;; E(d) = \frac{R_2 x C_2}{N}$$

Where N = a+b+c+d

To test this hypothesis we use the test statistic

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

the degrees of freedom for mxn contingency table is (m-1)x(n-1)

the degrees of freedom for 2x2 contingency table is (2-1)(2-1) = 1

This method is applied for all rxc contingency tables to get the expected frequencies.

The degrees of freedom for rxc contingency table is (r-1)x(c-1)

If the calculated value of χ^2 < table value of χ^2 at certain level of significance, then H_0 is accepted otherwise we do not accept H_0

The alternative formula for calculating χ^2 in 2x2 contingency table is

$$\chi^2 = \frac{(ad - bc)^2 N}{R_1 x R_2 x C_1 x C_2}$$

Example: Examine the following table showing the number of plants having certain characters, test the hypothesis that the flower colour is independent of the shape of leaf.

Flower colour	Shape of leaf		Totals
	Flat leaves	Curled leaves	
White flowers	99 (a)	36 (b)	$R_1 = 135$
Red flowers	20(c)	5 (d)	$R_2 = 25$
Totals	$C_1 = 119$	$C_2 = 41$	N = 160

Solution:

Null hypothesis H_0: attributes „flower colour‟ and „shape of leaf‟ are independent of each other.

Under H_0 the statistic is

$$\chi^2 = \frac{\sum_{i=1}^{n}(o_i - e_i)^2}{e_i}$$

where o_i = observed frequency
and e_i = expected frequency

Expected frequencies are calculated as follows.

$E(a) = \frac{R_1 * C_1}{N} = \frac{135 * 119}{160} = 100.40$ where R_1 and R_2 = Row totals

$E(b) = \frac{R_1 * C_2}{N} = \frac{135 * 41}{160} = 34.59$ C_1 and C_2 = column totals

$E(c) = \frac{R_2 * C_1}{N} = \frac{25 * 119}{160} = 18.59$ N= Grand totals

$E(d) = \frac{R_2 * C_2}{N} = \frac{25 * 41}{160} = 6.406$

o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2$	$\frac{(o_i - e_i)2}{e_i}$
99	100.40	-1.4	1.96	0.02
36	34.59	1.41	1.99	0.06
20	18.59	1.41	1.99	0.11
5	6.41	-1.41	1.99	0.31
				$\frac{\sum_{i=1}^{n}(o_i - e_i)^2}{e_i} = 0.49$

Calculated value of $\chi^2 = \frac{\sum_{i=1}^{n}(o_i - e_i)^2}{e_i} = 0.49$

Direct Method:

Statistic: $\chi^2 = \frac{N(ad - bc)^2}{R_1 R_2 C_1 C_2}$

here a = 99, b = 36, c = 20 and d = 5 and N = 160

$\chi^2 = \frac{160(99 * 5 - 36 * 20)^2}{}$

135 * 25 *119 * 41

$$= \frac{160 * 50625}{164666.25}$$

$$= \frac{18100000}{16466625} = 0.49$$

Calculated value of χ^2 = 0.40

Table value of χ^2 for (2-1) (2-1) = 1 d.f. is 3.84

Calculated value of χ^2 < Table value of χ^2 at 5% LOS for 1 d.f. , Null hypothesis is accepted and hence we conclude that two characters, flower colour and shape of leaf are independent of each other.

Yates correction for continuity in a 2x2 contingency table:

In a 2x2 contingency table, the number of d.f. is (2-1)(2-1) = 1. If any one of Expected cell frequency is less than 5, then we use of pooling method for χ^2 –test results with `0‟ d.f. (since 1 d.f. is lost in pooling) which is meaningless. In this case we apply a correction due to Yates, which is usually known a Yates correction for continuity.

Yates correction consists of the following steps; (1) add 0.5 to the cell frequency which is the least, (2) adjust the remaining cell frequencies in such a way that the row and column totals are not changed. It can be shown that this correction will result in the formula.

$$\chi^2 \text{(corrected)} = \frac{N\left[|ad - bc| - \frac{N}{2}\right]^2}{R_1 R_2 C_1 C_2}$$

Example: The following data are observed for hybrids of Datura.
Flowers violet, fruits prickly =47

Flowers violet, fruits smooth = 12
Flowers white, fruits prickly = 21

Flowers white, fruits smooth = 3. Using chi-square test, find the association between colour of flowers and character of fruits.

Sol: H_0: The two attributes colour of flowers and fruits are independent.

We cannot use Yate‟s correction for continuity based on observed values. If only expected frequency less than 5, we use Yates‟s correction for continuity.

The test statistic is

$$\chi^2_{(\text{corrected})} = \frac{N\left[\left|ad - bc\right| - \frac{N}{2}\right]^2}{R_1 R_2 C_1 C_2}$$

	Flowers Violet	Flowers white	Total
Fruits Prickly	47(48.34)	21(19.66)	68
Fruits smooth	12(10.66)	3(4.34)	15
Total	59	24	83

The figures in the brackets are the expected frequencies

$$\chi^2_{(corrected)} = \frac{83\left[\left|(47 * 3) - (21*12)\right| - \frac{83}{2}\right]^2}{68 *15 * 59 * 24}$$

$$= \frac{83\left[\left|141 - 252\right| - 41.5\right]^2}{68 *15 * 59 * 24}$$

$$= \frac{400910.75}{1444320} = 0.28$$

Calculated value of χ^2 = 0.28

Table value of χ^2 for (2-1) (2-1) = 1 d.f. is 3.84

Calculated value of χ^2 < table value of χ^2 , H_0 is accepted and hence we conclude that colour of flowers and character of fruits are not associated