NARSIMHA REDDY ENGINEERING COLLEGE
UGC AUTONOMOUS INSTITUTION
Maisammaguda (V), Kompally - 500100, Secunderabad, Telangana State, India
UGC - Autonomous Institute
Accredited by NBA & NAAC with 'A' Grade
Approved by AICTE
Permanently affiliated to JNTUH

UNIT-IV

Tests of Hypotheses (Large and Small Samples)

Testing of Hypothesis

Statistical Hypothesis: Hypothesis is a statement or assumption about the population which may or may not be true

Testing of hypothesis: It is used to testing the hypothesis about the parent population from which the samples are drawn.

Test of Significance: A very important aspect of the sampling theory is the study of the test of significance, which enables us to decide on the basis of the sample results, if

- The deviation between the observed sample statistics and the hypothesis parameter value (or)
- The deviation between two independent sample statistics is significant.

Null Hypothesis: A definite statement about the population parameter. Such hypothesis which is usually a hypothesis of no difference is called 'Null hypothesis' and is usually denoted by 'H_0'.

Alternative Hypothesis: Any hypothesis which is complementary to the null hypothesis is called 'Alternative hypothesis" and is usually denoted by 'H_1'.

Eg: If we want to test the null hypothesis that the population has a specified mean μ_0 (say)

i.e., H_1:μ≠ μ_0-----------------(i)

H_1:μ<μ_0-------------------(ii)

H_1:μ>μ_0--------------------(iii)

Then the alternative hypothesis in (i) is known as Two-tailed test and the alternatives in (ii)and (iii) are known as left and right tailed tests respectively.

Critical Region: The region of rejection of null hypothesis H_0 when H_0 is true is that region of the outcomes at where H_0 is rejected. If the sample points falls in that region is called the 'critical region', size of the critical region is α.

Type-I error:

P(rejecting H_0 /H_0 is true) i.e., when H_0 is true it is to be accepted but it is a rejected. Therefore there is a error

Type-II error: P(Accepting H_0/H_0 is false) i.e., when H_0 is false it is to be rejected but it is accepted. Therefore there is a error

One tailed and two tailed tests: If the alternative hypothesis is of the type (< or >) and the entire critical region lies in the normal probability curve on one side then it is said to be one tailed tests (OTT)

Again the one tailed test is two types (i) Right one tailed test (ii) Left one tailed test

If the alternative hypothesis is of the type ($\neq$) and the

critical region lies in the normal probability curve on both sides then it is said to be two tailed tests (TTT)

Level of significance (LOS): The probability of committing Type-I error is known as the level of significance which is denoted by 'α'. Usually LOS are 10% , 5% or 1%.

Degrees of freedom: Suppose there are N observations and k conditions on these then the degrees of freedom is N-k. The degrees of freedom is used in small sample tests.

Procedure for testing of hypothesis:

Step (1): Set up Null hypothesis (H_0)

Step (2): Set up Alternative hypothesis (H_1) Which enables us to apply one tailed test/ Two tailed test.

Step (3): Choose Level of significance (LOS) α

Step (4): Under the null hypothesis H_0, the test statistic Z=$\frac{t-E(t)}{S.E\ of(t)} \sim N(0,1)$ where 't ' is a statistic

Step (5): Conclusion: If calculated Z < (tabulated) Z_α at α % LOS then accept null hypothesis otherwise reject null hypothesis.

The rejection rule for H_0: x=μ (or) μ=μ_0 is given below.

Table: Critical value of Z when n≥30			
Level of significance	1%	5%	10%
Two-Tailed test	2.58	1.96	1.645
Right-Tailed test	2.33	1.645	1.28
Left-Tailed test	-2.33	-1.645	-1.28

Test of significance for a single mean: Working Rule

Step (1): Null hypothesis:$H_{0:}$μ=μ_0

Step (2): Alternative hypothesis: H_1:μ≠ $\frac{\mu_0}{}$ H_1:μ<μ_0 / H_1:μ>μ_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: We have the following two cases.

Case (1): When the S.D (σ) of population is known. Then the test statistic is Z=$\frac{\mu}{S.E()}\sim N(0,1)$

Where S.E($\overline{x}$) =$\frac{\sigma}{\sqrt{n}}$

Where σ = standard deviation of population

n=Sample size.

Case (2): When the S.D (σ) of population is unknown. The test statistic is Z=$\frac{\mu}{S.E()}$

Where S.E ($\overline{x}$) =$\frac{s}{\sqrt{n}}$

Where s = standard deviation of sample

n=Sample size.

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38.Also calculate 95% confidence interval for the population?

Given n=400, $\bar{x}$ =40, μ=38, $\sigma = 10$

Step (1): Null hypothesis:$H_{0:}$μ=38

Step (2): Alternative hypothesis: H_1:μ≠ 38

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

Step (4): Test statistic: When the S.D (σ) of population is known. Then the test statistic is

Z= $\frac{\mu}{S.E()}$ Where S.E ($\overline{x}$) =$\frac{\sigma}{\sqrt{n}}$

=4

Step (5): Conclusion: Z_{cal} =4, Z_{tab} =1.96

If Z_{cal}> Z_{tab} at 5% LOS. So we reject H_0.

95% confidence interval is ($\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$) =(39.02,40.98)

Test of equality of Two means: Let $\overline{x_1}, \overline{x_2}$be the sample means of two independent random samples sizes n_1 and n_2 drawn from two populations having the means μ_1 and μ_2 and standard deviation σ_1 andσ_2. To test whether the two population means are equal .

Step (1): Null hypothesis: $H_0: \mu_1 = \mu_2$

Step (2): Alternative hypothesis:

$$H_1 : \mu_1 \neq \mu_2 \,/\, H_1 : \mu_1 \leq_2 \,/\, H_1 : \mu_1 \geq \mu_2$$

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{\sigma^2{}_2}{n2}}}$

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If $Z_{cal} < Z_{tab}$ accept H_0. Otherwise reject H_0.

Problem:

The mean of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from same population of s.d 2.5 inches?

Sol : Given n_1=1000,n_2=2000 and $\overline{x_1}$ = 67.5 $\overline{x_2}$ = 68 population S.D σ=2.5

Step (1): Null hypothesis:$H_{0:}\mu_1 = \mu_2$

Step (2): Alternative hypothesis:

$H_1: \mu_1 \neq \mu_2$

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{\sigma^2{}_2}{n2}}}$ =5.16

Step (5): Conclusion: Z_{cal}=5.16 Z_{tab} =1.96

If $Z_{cal} > Z_{tab}$ then we reject our H_0.

∴ The samples have not been drawn from same population of S.D 2.5 inches.

Test of significance of single proportion: Suppose a large random sample of size n has a sample proportion p of members possessing a certain attribute. To test the hypothesis that the proportion P in the population has a specified value p_0.

Step (1): Null hypothesis:$H_{0:}$p=p_0

Step (2): Alternative hypothesis: H_1:p≠ p_0 H_1:p<p_0 / H_1:p>p_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$

Where p=sample proportion=x/n

P=population proportion, Q=1-P

N=sample size

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject $\square_0$.

Test of significance of single proportion: Suppose a large random sample of size n has a sample proportion p of members possessing a certain attribute. To test the hypothesis that the proportion P in the population has a specified value p_0.

Step (1): Null hypothesis:$H_{0:}$p=p_0

Step (2): Alternative hypothesis: H_1:p≠ $\frac{p_0}{}$ H_1:p<p_0 / H_1:p>p_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$

Where p=sample proportion=x/n

P=population proportion, Q=1-P

N=sample size

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem: In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% LOS?

Sol : Given n=1000

P=sample proportion of rice eaters=540/1000=0.54

P=population proportion of rice eaters=1/2=0.5 Q=1-P=0.5

Step (1): Null hypothesis:H_0: both rice and wheat are equally popular in the state.i.,e P=0.5

Step (2): Alternative hypothesis: H_1:p≠ 0.5

Step (3): Level of significance: 1% =2.58

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$ =2.532

Step (5): Conclusion: Z_{cal} = 2.532, Z_{tab} =2.58.

If Z_{cal}<Z_{tab} at 1% LOS then we accept H_0.

TESTING OF HYPOTHESIS (SMALL SAMPLES)

In the earlier chapter, we discussed certain tests which are valid only for large samples and hence based on the theory of the Normal distribution and literature on Tests of hypothesis like null hypothesis, Alternative hypothesis, Simple and composite hypothesis, Acceptance and rejection (critical) regions, level of significance, one tailed and two tailed tests etc.

In this chapter, apart from all the above said topics, in addition we need to concentrate on 'degrees of freedom'.

Degrees of freedom: It is very clear that in a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of the sample varies since it depends either on the experimenter or on the resources available. Moreover, the test statistic involves the estimated value of the parameter which depends on the number of observations. Hence the sample size plays an important role in testing of hypothesis and is taken care of by degrees of freedom.

Definition: The number of independent observations in a set is called degrees of freedom. It is denoted by ν (read as Nu). In general, the number of degrees of freedom is equal to the total number of observations less than the number of independent constraints imposed on the observations. i.e. in a set of n observations, if k is the number of independent constraints then $\nu = n - k$.

Before going to discuss the tests of significance under small samples, we need some knowledge about exact sampling distributions: t- distribution (or Student's t- distribution),

F- distribution and χ^2 - distribution (or Chi-Square distribution).

t- distribution: It is discovered by W.S.Gosset in 1908. The statistician Gosset is better known by the pen name (pseudonym) 'student' and hence t- distribution is called student's t- distribution.

In practice, the standard deviation σ is not known and in such a situation the only alternative left is to use S, the sample estimate of standard deviation σ. Thus, the variate $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ is approximately normal provided n is sufficiently large. If n is not sufficiently large (small) the variate $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ is distributed as t and hence, $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$ where $S^2 = \frac{1}{(n-1)}\sum_i (x_i - \bar{x})^2$.

Properties of t- distribution:

(1) The shape of t- distribution is bell-shaped and is symmetrical about mean.

(2) The curve of t- distribution is asymptotic to the horizontal axis.

(3) It is symmetrical about the line t = 0.

(4) The form of the probability curve varies with degrees of freedom.

(5) It is uni modal with mean = median = mode.

(6) The mean of t- distribution is zero and variance depends upon the parameter ν, is called the degrees of freedom.

(7) The t- distribution with ν degrees of freedom approaches standard normal distribution as $\nu \to \infty$, ν being a parameter.

The t- distribution is extensively used in hypothesis about one mean or single mean, or about equality of two means or difference of means when σ is known.

Some applications of t- distribution are:

(1). To test the significance of the difference between two sample means or to compare two samples.

(2). To test the significance of an observed sample correlation coefficient and sample correlation coefficient.

(3). To test the significance of difference between two sample means or to compare two samples.

Assumptions about t- test: t- test is based on the following five assumptions.

(1). The random sample has been drawn from a population.

(2). All the observations in the sample are independent.

(3). The sample size is not large. (One should note that at least five observations are desirable for applying a t- test.)

(4). The assumed value μ_0 of the population mean is the correct value.

(5). The sample values are correctly taken and recorded.

(6).The population standard deviation σ is unknown

In case the above assumptions do not hold good, the reliability of the test decreases

Problem: **The life expectancy of people in the year 1970 in Brazil is expected to be 50 years. A survey was conducted in 11 regions of Brazil and the data obtained are given below. Do the data confirm the expected view?**

Life expectancy (in years): 54.2, 50.4, 44.2, 49.7, 55.4, 57.0, 58.2, 56.6, 61.9, 57.5, 53.4.

Solution:

Null hypothesis, H_0: $\mu = 50$.

Alternative hypothesis, H_1: $\mu \neq 50$ (Two-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$, $d_i = x_i - y_i$ and $S^2 = \frac{1}{(n-1)} \sum_i (x_i - \bar{x})^2$.

Calculation of $\bar{x}$ and S^2 :

$$\bar{x} = \frac{54.2 + 50.4 + 44.2 + 49.7 + 55.4 + 57.0 + 58.2 + 56.6 + 61.9 + 57.5 + 53.4}{11} = \frac{598.5}{11} = 54.41.$$

and

$$S^2 = \frac{(54.2 - 54.41)^2 + (50.4 - 54.41)^2 + (44.2 - 54.41)^2 + ... + (53.4 - 54.41)^2}{10} = \frac{236.07}{10} = 23.607$$

$\Rightarrow S = 4.853.$

Therefore, the value of test statistic is

$$t = \frac{54.41 - 50}{4.859 / \sqrt{11}} = 3.01 .$$

t- critical value or table value at 5% level of significance and 10 degrees of freedom is 2.228 (from t- tables).

Since t- calculated value is greater than t- critical value, we reject the null hypothesis, H_0.

i.e. we accept H_1. It means that the life expectance more than 50 years.

Problem: **Mean life time of computers manufactured by a company is 1120 hours. (a) Test the hypothesis that mean lifetime of computers has not changed if a sample of 8 computers has a mean lifetime of 1070 hours with a standard deviation of 125 hours. (b) Is there decrease in mean lifetime? Use (i) 0.05 and (ii) 0.01 level of significance.**

Solution:

(a) Null hypothesis, H_0: $\mu = 1120$.

Alternative hypothesis, H_1: $\mu \neq 1120$ (Two-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$ and $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, a sample variance.

Calculation: We are given n = 8, $\bar{x} = 1070$, $s = 125$

Therefore, the value of test statistic is

$$t = \frac{1070 - 1120}{125/\sqrt{7}} = -1.05$$
$$or\ \ |t| = 1.05.$$

(i) t- critical value or table value at 5% level of significance with 7 degrees of freedom is 2.365 Since t- calculated value is less than t- critical value, we accept the null hypothesis, H_0.

i.e the sample has been come from the population whose mean life-time of computers is 1120 hours.

(ii) t- critical value or table value at 1% level of significance with 7 degrees of freedom is 3.499 and we accept null hypothesis H_0.

(b) **Null hypothesis**, H_0: $\mu = 1120$.

Alternative hypothesis, H_1: $\mu < 1120$ (left-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$ and $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, a sample variance.

Calculation: We are given n = 8, $\bar{x} = 1070$, $s = 125$

Therefore, the value of test statistic is

$$t = \frac{1070 - 1120}{125/\sqrt{7}} = -1.05$$
$$or\ \ t = -1.05$$

(i) t- critical value at 5% level of significance with 7 degrees of freedom for left tailed test is -1.895. Since t- calculated value is greater than t- critical value, we accept the null hypothesis, H_0.

i.e .the sample has been come from the population whose mean life-time of computers is 1120 hours.

(ii) t- critical value at 1% level of significance with 7 degrees of freedom is -2.998.

Since t- calculated value is greater than t- critical value, we accept the null hypothesis. i.e. it indicates no decrease in mean lifetime at either of the level of significances.

(These are the applications of t- test for single mean.)

Confidence limits: For example, 95% confidence limits for the population mean μ are $\bar{x} \pm t_\alpha \frac{S}{\sqrt{n}}$ or $\bar{x} \pm t_\alpha \frac{s}{\sqrt{n-1}}$ where $\alpha = 0.025$ for two-tailed test and $\alpha = 5\%$ for one-tailed test, as mentioned in the above example. i.e. For two-tailed test at α los, the value of is taken for $\alpha/2$ from statistical tables of t.

t-test for difference of means:

Assumptions:(i) parent populations , from which the samples have been drawn are normally distributed

(ii) The population variances are equal and unknown

(iii) The two samples are random and independent of each other

Suppose we want to test if two independent samples xi (i=1,2..n1) and yj(j=1,2..n2) of sizes n1 and n2 have been drawn from two normal populations with μ_x and μ_y respectively.

Under the null hypothesis H_0, $\mu_x = \mu_y$, H_1: $\mu_x \neq \mu_y$,

then test statistic

$$t = \frac{\bar{x} - \bar{y}}{S / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{\sum_i x_i}{n}$, $\bar{y} = \frac{\sum_i y_i}{n}$ and $S^2 = \frac{1}{(n_1+n_2-2)}\left[\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2\right]$, a combined sample mean square.

Problem: A group of 5 patients with medicine A weigh 42, 39, 48, 60, and 41 kilograms. Another group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kilograms. Do you agree with the claim that medicine B increase the weigh significantly?

Solution: Null hypothesis, H_0: There is no significant difference between the medicines A and B with reference to their effect on increase in weight. i.e. $\mu_x = \mu_y$.

Alternative hypothesis, H_1: $\mu_x < \mu_y$ (left-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{\sum_i x_i}{n}$ and $S^2 = \frac{1}{(n_1 + n_2 - 2)}\left[\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2\right]$, a combined sample mean square.

Calculation of $\bar{x}$ and S^2:

We are given $n_1 = 5$, $n_2 = 7$.

From the given data,

$$\bar{x} = \frac{42+39+48+60+41}{5} = \frac{230}{5} = 46,$$

$$\bar{y} = \frac{38+42+56+64+68+69+62}{7} = \frac{399}{7} = 57$$

and

$$\sum_{i=1}^{5}(x_i - 46)^2 = (42-46)^2 + (39-46)^2 + \ldots + (41-46)^2 = 290$$

$$\sum_{i=1}^{7}(y_i - 57)^2 = (38-57)^2 + (42-57)^2 + \ldots + (62-57)^2 = 926$$

Hence,

$$S^2 = \frac{290+926}{5+7-2} = 121.6$$

$$\Rightarrow S = 11.03.$$

Therefore, the value of test statistic is

$$t = \frac{46-57}{11.03\sqrt{\frac{1}{5}+\frac{1}{7}}} = -\frac{11}{6.46} = -1.7$$

$$\therefore t = -1.7$$

t- Critical value at 5% los with 10 degrees of freedom for right tailed test is -1.812.

Clearly, t- calculated value is greater than t- critical value at 5% los, we accept the null hypothesis. i.e. the medicines A and B do not differ significantly with reference to their effect on increase in weight.

(This is an application of Test for difference of means.)

Problem: **Memory capacity of 10 students was tested before and after training. State whether the training was effective or not from the following scores:**

Before training:	**12**	**14**	**11**	**8**	**7**	**10**	**3**	**0**	**5**	**6**
After training:	**15**	**16**	**10**	**7**	**5**	**12**	**10**	**2**	**3**	**8**

Solution:

Null hypothesis, H_0: There is no significant effect of the training on memory capacity of the students. i.e. $\mu_1 = \mu_2$.

Alternative hypothesis, H_1: Memory capacity of the students has been increased or improved after training. i.e. $\mu_1 < \mu_2$ (left-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

Where $\bar{d} = \frac{\sum_i d_i}{n}$, $d_i = x_i - y_i$ and $S^2 = \frac{1}{(n-1)} \sum_i (d_i - \bar{d})^2$.

Calculation of $\bar{d}$ and S^2:

Let memory capacity before training and after training be x and y respectively.

$$\bar{d} = \frac{(12-15)+(14-16)+(11-10)+(8-7)+(7-5)+(10-12)+(3-10)+(0-2)+(5-3)+(6-8)}{10}$$

$$= -\frac{12}{10} = -1.2$$

and

$$S^2 = \frac{(-3-(-1.2))^2 + (-2-(-1.2))^2 + (1-(-1.2))^2 + \ldots + (-2-(-1.2))^2}{9} = 7.73$$

$$\Rightarrow S = 2.78$$

Therefore, the value of test statistic is

$$t = \frac{-1.2}{2.78/\sqrt{10}} = -1.365.$$

t- Critical value at 5% los with 9 degrees of freedom for left-tailed test is -1.833.

Since t- calculated value is greater than t- critical value at 5% los with 9 degrees of freedom for left tailed test, we accept the null hypothesis, H_0 i.e. we conclude that there is no change in memory capacity after the training programme (or) there is no use of training programme.

F-distribution :- "The ratio of two sample variances is distributed of F." F-distribution was worked out by G.W. Snedecor and as a mark of respect for Sir R.A.Fisher (Father of modern statistics). Who was defined a statistics Z which is based upon the ratio of two –sample variances initially and hence it is denoted by F. (The first letter of Fisher).

Let s_1^2 be the sample variance of an independent sample of size n_1 drawn from a normal population N (μ_1, σ_1^2). Similarly , let S_2^2 be the sample variance in an independent sample of size n2 drawn from another normal population N (μ_2, σ_2^2).Thus s_1^2 and S_2^2 are the variances of two random samples of sizes n_1 and n_2 respectively drawn from two normal populations. In order to determine whether the two samples came from two populations having equal variances of the two independent random samples defined by

F= $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \dfrac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$

Which is an F-distribution with v1=n1-1 and v2=n2-1 degrees of freedom.

Properties of F-distribution:

(1) F-distribution curve extends on abscissa from 0 to ∞.

(2) It is an unimodal curve and its mode lies on the point

F= $\dfrac{k_2(k_1-2)}{k_1(k_2+2)}$ or $\dfrac{v_2(v_1-2)}{v_1(v_2+2)}$ which is always less than unity

(3) F-distribution curve is a positive skew curve. Generally, the F-distribution curve is highly positive skewed where v2 is small

(4) The mean and variance are defined when v2 ≥ 3 and v2≥5 respectively.

(5) There exists a very useful relation for interchange of degrees of freedom v1 and v2 i.e

$$F_{1-\alpha}(v_1, v_2) = \frac{1}{F_\alpha(v_2, v_1)}$$

(6) The moment generating function of F-distribution does not exist.

F-test is used to

(1)Test the hypothesis about the equality of two population variances.

(2) test the hypothesis about the equality of two or more population means.

F-test for equality of two population variances: Suppose we want to test whether two independent samples xi (i=1,2..n1) and yj(j=1,2..n2) of sizes n1 and n2 have been drawn from two normal populations with the same variance or not then

Null hypotheses, $H_0 : \sigma_x^2 = \sigma_y^2$, $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Under the null hypothesis, H_0, the test statistics is:

$$F = \frac{s_x^2}{s_y^2} \sim F_{(v_1,v_2)} \quad \text{(OR)} \quad F = \frac{s_y^2}{s_x^2} \sim F_{(v_2,v_1)}$$

When $s_x^2 > s_y^2$ OR $s_y^2 > s_x^2$ respectively

Where $s_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$ with $\bar{x} = \frac{\sum_{i-1}^{n_1} x_i}{n_1}$

And $s_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$ with $\bar{y} = \frac{\sum_{i-1}^{n} y_i}{n_2}$

Besides t-test , we can also apply a F-test for testing equality of two population means. F-distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression coefficient etc.,

As a matter of fact , F-test is the backbone of analysis of variance(ANOVA)

Note: (1) F determines whether the ratio of two sample variances s1 and s2 is too small or too large.

(2) when F is close to 1, the two sample variances s1 and s2 are likely same

(3) F-distribution also known as variance ratio distribution

(4) Dividing S_1^2 and S_2^2 by their corresponding population variances standardizes the sample variance, and hence on the average both numerator and denominator approach. Therefore, its customer, to take the large sample variance as the numerator.

(5) F-distribution depends not only on the two parameters, V_1 and V_2 but also on the order in which they are slated.

Problem: Life expectancy in 9 regions of Brazil in 1900 and in 11 regions of Brazil in 1970 was as given in the table below:

(Source: The review of income and wealth, June 1983)

Regions / Life Expectancy	1	2	3	4	5	6	7	8	9	10	11
1900	42.7	43.7	34.0	39.2	46.1	48.7	49.4	45.9	55.3	-	-
1970	54.2	50.4	44.2	49.7	55.4	57.0	58.2	56.6	61.9	57.5	53.4

It is desired to confirm, whether the variation in life expectancy in various reigns in 1900 and in 1970 in same or not.

Solution: Let the populations in 1900 and in 1970 be considered as $N(\mu_1,\, \sigma_1^2)$ *and* $N(\mu_1,\, \sigma_1^2)$ respectively.

Null hypotheses, H_0 : The variation of life expectancy in various regions in 1900 and in 1970 is same. $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Under the null hypothesis, H_0, the test statistics is :

$$F = \frac{s_1^2}{s_2^2} \sim F_{(v_1, v_2)} \quad \text{(OR)} \quad F = \frac{s_2^2}{s_1^2} \sim F_{(v_2, v_1)}$$

When $s_1^2 > s_2^2$ OR $s_2^2 > s_1^2$ respectively

$$\text{Where } s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2 \text{ with } \bar{x} = \frac{\sum_{i-1}^{n_1} x_i}{n_1}$$

$$\text{And } s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(y_i - \bar{y}\right)^2 \text{ with } \bar{y} = \frac{\sum_{i-1}^{n} y_i}{n_2}$$

Calculation: $\bar{x} = \dfrac{405}{9} = 45,\ \bar{y} = \dfrac{598.5}{11} = 5441$(approximate)

$$\sum_{i=1}^{9}\left(x_i - 45\right)^2 = 5.29+1.69+121+33.64+1.21+13.69+0.9+106.9$$

= 288.51+19.36=302.87

$$\Rightarrow s_1^2 = \frac{302.87}{8} = 37.85$$

$$\sum_{i=1}^{11}\left(y_i - 54.41\right)^2 = 0.04+16+104.04+22.09+1+6.76+14.44+4.84+56.25+9.61+1 = 236.07$$

$$\Rightarrow s_2^2 = \frac{236.07}{10} = 23.607$$

Since $s_1^2 > s_2^2$, the value of test statistics is:

$$F = \frac{37.85}{23.607} = 1.603$$

The table value of F at 5% los with (8, 10) degrees of freedom for two tailed test is 3.85 (From F-tables).

Since F-Calculated value is less than f-tabulated value, we accept H_0. i.e. The sample data confirms the equality of variances in 1900 and 1970 in various regions of brazil or $\sigma_1^2 = \sigma_2^2$.

Practice.

Problem: The house-hold net expenditure on health care in south and north India, in two samples of households, expressed as percentage of total income is shown the following table:

South:	**15.0**	**8.0**	**3.8**	**6.4**	**27.4**	**19.0**	**35.3**	**13.6**	
North:	**18.8**	**23.1**	**10.3**	**8.0**	**18.0**	**10.2**	**15.2**	**190.0**	**20.2**

Test the equality of variances of households net expenditure on health care in south and north India.

Problem: **The time taken by workers in performing a job by method I and Method II is given below.**

Method I	**20**	**16**	**26**	**27**	**23**	**22**	**-**
Method II	**27**	**33**	**42**	**35**	**32**	**34**	**38**

Do the data show that the variances of time distribution of population from which these samples are drawn do not differ significantly?

Solution:

Null hypothesis(H_0 **)**: There is no significant difference between the variances of time distribution of populations. i.e. $\sigma_1^2 = \sigma_2^2$.

Alternative hypothesis(H_1): $\sigma_1^2 \neq \sigma_2^2$ (Two-tailed test)

Level of significance : Choose $\alpha = 5\% = 0.05$

Under the null hypothesis, H_0,

the test statistics is :

$$F = \frac{s_1^2}{s_2^2} \sim F_{(v_1,v_2)} \quad \text{(OR)} \quad F = \frac{s_2^2}{s_1^2} \sim F_{(v_2,v_1)}$$

When $s_1^2 > s_2^2$ OR $s_2^2 > s_1^2$ respectively

Calculation: we are given $n_1 = 6$, $n_2 = 7$

$$\bar{x} = \frac{134}{6} = 22.3, \ \bar{y} = \frac{241}{7} = 34.4$$

$$\sum_{i=1}^{6} (x_i - 22.3)^2 = 81.34, \ \sum_{i=1}^{7} (y_i - 34.4)^2 = 133.72$$

$$\therefore s_1^2 = \frac{81.34}{5} = 16.26 \text{ and } s_2^2 = \frac{133.72}{6} = 22.29$$

The value of test statistics is

$$F = \frac{22.29}{16.26} = 1.3699 \cong 1.37$$

F-critical value at 5% los with (5, 6) degrees of freedom for two tailed test is 4.39 (From F-tables)

Since F-Calculated value is less than F-tabulated value t 5% los, we accept H_0. i.e. there is no significant deference between the variances of the time distribution by the workers.

Problem: The nicotine contents in milligrams in two samples of tobacco were found to be as follows:

Sample A	24	27	26	21	25	-
Sample B	27	30	28	31	22	36

Can it be said that the two samples have come from the same normal population?

Solution

Since the two samples are independent, the sample sizes are small ($n < 30$), and the populations are assumed to be normal, we use the Student's two-sample t-test (assuming equal population variances).

Step 1: State the Hypotheses

Null Hypothesis:

$$H_0: \mu_1 = \mu_2$$

(The two population means are equal.)

Alternative Hypothesis:

$$H_1: \mu_1 \neq \mu_2$$

(The two population means are not equal.)

Step 2: Calculate Sample Means

Sample A

Values
24
27
26
21
25

$$n_1 = 5$$

$$\sum X_1 = 24 + 27 + 26 + 21 + 25 = 123$$

$$\bar{X}_1 = \frac{123}{5} = 24.6$$

Sample B

Values
27
30
28
31
22
36

$$n_2 = 6$$

$$\sum X_2 = 27 + 30 + 28 + 31 + 22 + 36 = 174$$

$$\bar{X}_2 = \frac{174}{6} = 29$$

Step 3: Calculate Sample Variances

Sample A

X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$
24	-0.6	0.36
27	2.4	5.76
26	1.4	1.96
21	-3.6	12.96
25	0.4	0.16

$$\sum (X_1 - \bar{X}_1)^2 = 21.20$$

$$s_1^2 = \frac{21.20}{5-1} = 5.30$$

Sample B

X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
27	-2	4
30	1	1
28	-1	1
31	2	4
22	-7	49
36	7	49

$$\sum(X_2 - \bar{X}_2)^2 = 108$$

$$s_2^2 = \frac{108}{6-1} = 21.60$$

Step 4: Calculate the Pooled Variance
The pooled variance is

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Substitute the values:

$$S_p^2 = \frac{4(5.30) + 5(21.60)}{5 + 6 - 2}$$

$$= \frac{21.2 + 108}{9}$$

$$= \frac{129.2}{9}$$

$$= 14.36$$

Therefore,

$$S_p = \sqrt{14.36} = 3.79$$

Step 5: Calculate the Test Statistic
The t-statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Substitute the values:

$$t = \frac{24.6 - 29}{3.79\sqrt{\frac{1}{5} + \frac{1}{6}}}$$

$$= \frac{-4.4}{3.79\sqrt{\frac{11}{30}}}$$

$$= \frac{-4.4}{3.79(0.6055)}$$

$$= \frac{-4.4}{2.30}$$

$$\boxed{t = -1.91}$$

Step 6: Determine the Critical Value
Degrees of freedom:

$$df = n_1 + n_2 - 2 = 5 + 6 - 2 = 9$$

At the 5% significance level (two-tailed),

$$t_{0.025,9} = 2.262$$

Step 7: Decision
Calculated value:

$$|\, t \,| = 1.91$$

Critical value:

$$2.262$$

Since

$$1.91 < 2.262$$

we fail to reject the null hypothesis.

Conclusion
There is no significant difference between the mean nicotine contents of the two samples at the 5% significance level.
Hence, the data support the assumption that the two samples have come from the same normal population.
Final Answer

- Sample A mean:

$$\boxed{\bar{X}_1 = 24.6}$$

- Sample B mean:

$$\boxed{\bar{X}_2 = 29.0}$$

- Pooled variance:

$$\boxed{S_p^2 = 14.36}$$

- Calculated t-value:

$$\boxed{t = -1.91}$$

- Critical t-value:

$$\boxed{\pm 2.262}$$

Decision
Since

$$|\,t\,| = 1.91 < 2.262,$$

Accept (fail to reject) the null hypothesis.

Problem
A random sample of 10 boys had the following IQ scores:

$$\mathbf{70, 120, 110, 101, 83, 88, 95, 98, 107, 100}$$

Test whether these data support the assumption that the population mean IQ is 100.
Take the 5% level of significance.

Solution
Since the sample size is less than 30 and the population standard deviation is unknown, we use a one-sample t-test.
Step 1: State the Hypotheses

- Null Hypothesis:

$$H_0: \mu = 100$$

- Alternative Hypothesis:

$$H_1: \mu \neq 100$$

(Two-tailed test)
Step 2: Given Data
Sample size,

$$n = 10$$

Hypothesized population mean,

$$\mu = 100$$

Sample observations:

70, 120, 110, 101, 83, 88, 95, 98, 107, 100

Step 3: Calculate the Sample Mean

IQ (X)

70

120

110

101

83

88

95

98

107

100

$$\sum X = 972$$

Hence,

$$\bar{X} = \frac{\sum X}{n}$$

$$= \frac{972}{10}$$

$$\boxed{\bar{X} = 97.2}$$

Step 4: Calculate the Sample Standard Deviation

Construct the following table.

X	$X - \bar{X}$	$(X - \bar{X})^2$
70	−27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
83	−14.2	201.64
88	−9.2	84.64
95	−2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84

$$\sum (X - \bar{X})^2 = 1833.60$$

The sample variance is

$$s^2$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{6} = 1.934$$

$$\sigma$$

n

xbar=-0.01s^2 ~ 1.934

$$s^2 = \frac{\sum(X-\bar{X})^2}{n-1}$$
$$= \frac{1833.6}{9}$$
$$= 203.73$$

Therefore,

$$s = \sqrt{203.73}$$
$$\boxed{s = 14.27}$$

Step 5: Calculate the t-Test Statistic
The test statistic is

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

Substitute the values:

$$t = \frac{97.2-100}{14.27/\sqrt{10}}$$
$$= \frac{-2.8}{14.27/3.162}$$
$$= \frac{-2.8}{4.513}$$
$$\boxed{t = -0.62}$$

Step 6: Critical Value
Degrees of freedom:

$$df = n-1 = 9$$

For a two-tailed test at the 5% significance level,

$$t_{0.025,9} = 2.262$$

Critical region:
Reject H_0 if

$$|\, t \,| > 2.262$$

Since

$$|\, t \,| = 0.62 < 2.262$$

we fail to reject H_0.
Step 7: Conclusion
There is no significant evidence at the 5% level to conclude that the population mean IQ differs from 100.
Therefore, the sample data support the assumption that the population mean IQ is 100.
Final Answer

- Sample mean:

$$\boxed{\bar{X} = 97.2}$$

- Sample standard deviation:

$$\boxed{s = 14.27}$$

- Calculated t-value:

$$\boxed{t = -0.62}$$

- Critical value:

$$\boxed{\pm 2.262}$$

Decision
Since

$$|\ t\ | = 0.62 < 2.262,$$

Accept (fail to reject) the null hypothesis.