UNIT-4 TEST FOR HYPOTHESIS FOR SMALL SAMPLES

Test of hypothesis for single mean

Student t-test for single mean

Let $\bar{x}$ be a mean of the sample, n = size of the sample, S = Standard deviation of the sample, M = mean of the population,

then the test statistic is

$$t = \frac{\bar{x} - M}{S/\sqrt{n-1}}$$

$$t = \frac{\bar{x} - M}{S/\sqrt{n}} \quad \text{where } S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

The Standard Error for single mean is $\frac{S}{\sqrt{n}}$

Maximum Error for single mean

$$E_{max} = Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

confidence limits for M is

$$(\bar{x} - E_{max}, \bar{x} + E_{max})$$

$$= \left(\bar{x} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

procedure

Let a random sample of size n where $n < 30$, a sample mean $\bar{x}$. To test the hypothesis that the population mean M has a specified value M_0 when population standard deviation σ is not known. Set the null Hypothesis as H_0.

① $H_0 : M = M_0$

Alternative has three possibilities

② $H_1 : M \neq M_0$

$M < M_0$

$M > M_0$

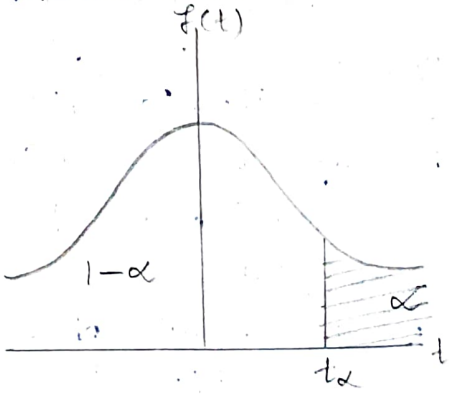
③ $t = \frac{\bar{x} - M}{\frac{S}{\sqrt{n-1}}}$ (or)

$$t = \frac{\bar{x} - M}{\frac{S}{\sqrt{n}}} \quad \text{where } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where s is the sample standard deviation follows t distribution with $n-1$ degrees of freedom.

Properties of t distribution

- The shape of t distribution is well-shaped which is similar to that of a normal distribution and is symmetrical about the mean.
- t distribution is also asymptotic to the t-axis that is the two tiles of the curve on both sides of $t = 0 \longrightarrow \infty$
- It is unimodal with mean = median = mode.



- The mean of standard Normal distribution and as well as t-distribution is 0.
- But the variance of t distribution depends on the degree of freedom. The selected values of t_α for various degrees of freedom can be obtained from the tables of t distribution where t_α denotes the area under t distribution is equal to 0 to its right.

Applications of the t distribution

- The t distribution has a wide number of applications in statistics. some of them are
  - → To test the significance of the sample mean when population variance is not given.
  - → To test the significance of the difference of sample means.
  - → To test the significance of an observed sample corelation coefficient and sample regression coefficient.

Sample important questions

- Write the suitable test statistic for test of significance of single mean in small samples.
- Write the procedure for test of significance of single mean in small samples.

1) A sample of 26 Bulbs gives a mean life of 990 hours with a standard deviation of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard?

Sol:- Given that
the sample size (n) = $26 < 30$ (Small sample)
The mean of the sample ($\bar{x}$) = 20 hours
$\mu_1 = 1000$ hours

(1) $H_0 : \mu = 1000$

(2) $H_1 : \mu < 1000$ (Left tailed test)

(3) LOS $\alpha = 5\%$.
$t = 0.05$ with 25 degrees of freedom $= 1.708$

degrees of freedom (v) $= n - 1$
$= 26 - 1$
$= 25$

(4) Test statistic $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}}$

$t = \dfrac{990 - 1000}{20/\sqrt{26-1}}$

$= \dfrac{-10}{20} \times \sqrt{25}$

$= 2.5$

(5) Conclusion $|t_{cal}| = 2.5 > t_{tab}$
We reject H_0

Therefore the given sample is not upto the standard.

2) A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 washers has found to have a thickness of 0.024 cm with standard deviation of 0.002 cm. Test the significance of the deviation.

Sol:- $\mu = 0.025$ cm
$\bar{x} = 0.024$ cm
$s = 0.002$ cm
$n = 10$

(1) $H_0 : \mu = 0.025$ cm

(2) $H_1 : \mu \neq 0.025$ (Two tailed test)

(3) LOS $\alpha = 5\% = 0.05$,

$\frac{\alpha}{2} = 0.025$.

$\nu = n - 1 = 10 - 1 = 9$

$t_{0.025}$ with 9 degrees of freedom $= 2.262$

(4) Test statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$

$$t = \frac{0.024 - 0.025}{0.002/\sqrt{9}}$$

$$= 1.5$$

(5) conclusion $t_{cal} < t_{tab}$

accept H_0

3) The average breaking strength of steel rods is specified to be 18.5 1000 pounds. To Test this sample of 14 rods were tested. The mean and standard deviations were obtained were 17.85, 1.955 1000 pounds respectively. Is the result of the experiment significant.

Sol:- $\mu = 18.5$

$\bar{x} = 17.85$

$s = 1.955$

$n = 14$.

(1) $H_0 : \mu = 18.5$

(2) $H_1 : \mu \neq 18.5$ (Two tailed test)

(3) LOS $\alpha = 5\% = 0.05$,

$\frac{\alpha}{2} = 0.025$

$\nu = n - 1 = 14 - 1 = 13$

$t_{0.025}$ with 13 degrees of freedom $= 2.160$

(4) Test statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$

$$t = \frac{17.85 - 18.5}{1.955/\sqrt{3}}$$

$$= -0.57$$

(5) conclusion $|t_{cal}| < t_{tab}$ accept H_0

4) A random sample of size 16 values from a normal population showed a mean of 53 and a sum of squares of deviations from the mean = 150. Can the sample be regarded as taken from the population having 56 as mean? Obtain 95% confidence limits of the mean of the population?

Sol:- $n = 16$

$\bar{x} = 53$

$\mu = 56$

(1) $H_0 : \mu = 56$

(2) $H_1 : \mu \neq 56$ (Two tailed test)

(3) LOS $\alpha = 5\% = 0.05$

$\frac{\alpha}{2} = 0.025$

$\nu = n - 1 = 16 - 1 = 15$

$t_{0.025}$ with 15 degrees of freedom $= 2.131$

(4) Test statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$

$\Sigma (x_i - \bar{x})^2 = 150$

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n} = \frac{150}{16}$$

$$= 9.38$$

$s = \sqrt{9.38} = 3.06$

$$t = \frac{53-56}{\sqrt{9.375}/\sqrt{15}}$$

$$= -3.79$$

(5) conclusion $|t_{cal}| > t_{tab}$

we reject H_0, accept H_1

sample cannot be regarded as taken from the population having 56 as mean.

Note :-

t-test

$t = \frac{\bar{x}-M}{S/\sqrt{n-1}}$ S is known

$t = \frac{\bar{x}-M}{S/\sqrt{n}}$ $S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

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
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
		1833.60

Small samples - $Z_{\alpha/2}$

Large samples - $t_{\alpha/2}$

(1) A random sample of 10 boys had the following IQ's.

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

i) Do this data support the assumption of a population mean IQ of 100.

ii) Find a reasonable range in which most of the mean IQ value of samples of 10 boys likely.

Soln i) Given that size of the Sample $(n) = 10$

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

$$= \frac{70+120+110+101+88+83+95+98+107+100}{10}$$

$$= 97.2.$$

$\mu = 100$

$$\sum_{i=1}^{100}(x_i-\bar{x})^2 = 1833.6$$

$$S^2 = \frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{10-1} = \frac{1833.6}{9}$$

$$S^2 = 203.73$$

$$S = \sqrt{203.73} = 14.27$$

(1) Null Hypothesis $H_0 : M = 100$

(2) Alternative Hypothesis $H_1 : M \neq 100$ (Two tail test)

(3) Level of significance

$\alpha = 5\%$. $\frac{\alpha}{2} = 0.025$

degrees of freedom is $n-1 = 9$

from table $t_{0.025}$ with 9 degrees of freedom is 2.262

(4) Test statistic $t = \frac{\bar{x}-M}{S/\sqrt{n}}$

$$= \frac{97.2-100}{14.27/\sqrt{10}}$$

$$= -0.62$$

(5) conclusion : $t_{cal} < t_{tab}$

accept H_0

ii) Confidence Interval

$$(\bar{x} - t_{\alpha/2} S/\sqrt{n} \;,\; \bar{x} + t_{\alpha/2} S/\sqrt{n})$$

$$= \left(97.2 - 2.262 \times \frac{14.27}{\sqrt{10}} \;,\; 97.2 + 2.262 \times \frac{14.27}{\sqrt{10}}\right)$$

$$= (87, 107.4)$$

---

2) The height of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches.

Is it reasonable to believe that the average height is greater than 64 inches.

Test at 5% Level of significance.

Sol:- Given that size of the sample

$$n = 10$$

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{70+67+62+68+61+68+70+64+64+66}{10}$$

$$= 66$$

x	$x - \bar{x}$	$(x-\bar{x})^2$
70	4	16
67	1	1
62	-4	16
68	2	4
61	-5	25
68	2	4
70	4	16
64	-2	4
64	-2	4
66	0	0
		90

$$S^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{10-1}$$

$$= \frac{90}{9} = 10$$

$$S^2 = 10$$

$$S = \sqrt{10}$$

(1) $H_0 : \mu = 64$

(2) $H_1 : \mu > 64$ (right tail test)

(3) Level of significance

$\alpha = 5\%$

degrees of freedom $= n - 1 = 9$

from table $t_{0.05}$ with 9 degrees of freedom is 1.833

(4) Test statistic $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$

$$= \frac{66 - 64}{\sqrt{10}/\sqrt{10}} = 2$$

(5) conclusion : $t_{cal} > t_{tab}$

reject H_0, accept H_1

The average height is greater than 64 is acceptable statement or believable statement.

3) A random sample from a companies very extensive files shows that the orders for a certain piece of machinery were filled, respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Test the claim on the average such orders are filled in 10.5 days. Use the level of significance $\alpha = 0.01$. To choose the alternative hypothesis so that rejection of Null Hypothesis $\mu = 10.5$ days $\Rightarrow$ that it takes longer than indicated.

Sol:- $\bar{x} = \frac{10+12+19+14+15+18+11+13}{8}$

$\bar{x} = 14$

x	$x - \bar{x}$	$(x-\bar{x})^2$
10	-4	16
12	-2	4
19	5	25
14	0	0
15	1	1
18	4	16
11	-3	9
13	-1	1
		72

$s^2 = \frac{\sum_{i=1}^{8} (x_i - \bar{x})^2}{8-1} = \frac{72}{7}$

$s^2 = 10.28$

$s = 3.20$

① $H_0 : \mu = 10.5$

② $H_1 : \mu > 10.5$ (right tail test)

③ Level of significance : $\alpha = 1\%$.

degrees of freedom $= 8 - 1 = 7$

from table $t_{0.01}$ with 7 degrees of freedom is 2.988

④ Test statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$= \frac{14 - 10.5}{3.20/\sqrt{8}}$

$= 3.093$

⑤ $t_{cal} > t_{tab}$, reject H_0 and accept H_1

4) A random sample of 8 envelops is taken from the letter box of post office and their weights in grams are found to be 12.1, 11.9, 12.4, 12.3, 11.5, 11.6, 12.1 and 12.4.

i) Does this sample indicate at 1% level that the average weight of envelops received at their post office is 12.35 grams?

ii) Find 95% confidence limits for the mean weight of the envelops received at that post office.

Sol:- Given that size of the sample $n = 8$

$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.1+11.9+12.4+12.3+11.5+11.6+12.1+12.4}{8}$

$= 12.03$

x	$x-\bar{x}$	$(x-\bar{x})^2$
12.1	0.07	0.0049
11.9	−0.13	0.0169
12.4	0.37	0.1369
12.3	0.27	0.0729
11.5	−0.53	0.2809
11.6	−0.43	0.1849
12.1	0.07	0.0049
12.4	0.37	0.1369
		0.84

$$S^2 = \frac{\sum_{i=1}^{8}(x_i-\bar{x})^2}{7}$$

$$= \frac{0.84}{7}$$

$$S^2 = 0.12$$

$$S = 0.346 \simeq 0.35$$

(1) $H_0 : M = 12.35$

(2) $H_1 : M \neq 12.35$ (two tail test)

(3) i) Level of significance

$\alpha = 1\% = 0.01$

degrees of freedom $= n-1$

$\Rightarrow 8-1 = 7$

$$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$

from table $t_{0.005}$ with 7 degrees of freedom is 3.499

(4) Test statistic $t = \dfrac{\bar{x}-M}{S/\sqrt{n}}$

$$= \frac{12.03-12.35}{0.35/\sqrt{8}}$$

$$= -2.58$$

(5) conclusion : $|t_{cal}| < t_{tab}$

accept H_0

ii)

(3) Level of significance

$\alpha = 5\% = 0.05$

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

from table $t_{0.025}$ with 7 degrees of freedom is 2.365

(5) conclusion : $|t_{cal}| > t_{tab}$

reject H_0, accept H_1

Confidence interval

$$(\bar{x} - t_{\alpha/2}\, S/\sqrt{n},\ \bar{x} + t_{\alpha/2}\, S/\sqrt{n})$$

$$= \left(12.03 - 2.365 \times \frac{0.35}{\sqrt{8}},\ 12.03 + 2.365 \times \frac{0.35}{\sqrt{8}}\right)$$

$$= (11.73, 12.32)$$

Test of significance for difference of two means

Let $\bar{x}$ and $\bar{y}$ be the means of two independent samples of sizes n_1 and n_2 ($n_1 < 30$) and ($n_2 < 30$) drawn from two normal populations having means M_1 and M_2. To test whether the two population means are equal.

Let the null hypothesis be

$H_0 : M_1 = M_2$

Then the alternative hypothesis be $H_1 : M_1 \neq M_2$

$M_1 < M_2$

$M_1 > M_2$

The test statistic $t = \dfrac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

Where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

(or)

$$\frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{j=1}^{n_2}(y_j - \bar{y})^2}{n_1 + n_2 - 2}$$

which follows the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The Standard Error of $\bar{x} - \bar{y}$ is $S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Maximum Error is $t_{\alpha/2} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

confidence Interval is,

$$\left((\bar{x} - \bar{y}) - t_{\alpha/2} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{x} - \bar{y}) + t_{\alpha/2} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

1) Samples of two types of electrical light bulbs were tested for length of life and following data were obtained.

	Type1	Type 2
Sample size	8	7
sample mean	1234 hrs	1036 hrs
Sample Standard Deviation	36 hrs	40 hrs

Is the difference in the means sufficient to warrant that type1 is superior to type2 regarding length of life

Sol:- Given that $n_1 = 8$

Mean of the first sample ($\bar{x}$)

$= 1234$ hours

Standard deviation of first sample (S_1) = 36 hours

$n_2 = 7$

Mean of the second sample ($\bar{y}$)

$= 1036$ hours

Standard deviation of second sample (S_2) = 40 hours

(1) Null Hypothesis $H_0: M_1 = M_2$

(2) Alternative Hypothesis

$H_1: M_1 > M_2$ (right tailed test)

(3) $\alpha = 0.05$

degrees of freedom $= n_1 + n_2 - 2$

$= 15 - 2$

$= 13$

from table $t_{0.05}$ with 13 degrees of freedom is 1.771

(4) Test Statistic $t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$= \frac{8(36)^2 + 7(40)^2}{8 + 7 - 2}$

$= 1659.08$

$S = \sqrt{1659.08}$

$t = \frac{1234 - 1036}{\sqrt{1659.08\left(\frac{1}{8} + \frac{1}{7}\right)}}$

$t = 9.39$

(5) Conclusion:

$t_{cal} > t_{tab}$, reject H_0

$\Rightarrow$ accept H_1

$\therefore$ type1 is not superior to type2 regarding length of life.

2) The IQ's of 16 students from one area of a city showed a mean of 107 with a standard deviation of 10, while the IQ's of 14 students from another area of a city showed a mean of 112 with a standard deviation of 8. Is there significant difference between the IQ's of the two groups at 5% level of significance.

Sol:- Given that $n_1 = 16$

$\bar{x} = 107$, $s_1 = 10$

$n_2 = 14$, $\bar{x} = 112$, $s_2 = 8$

(1) Null Hypothesis $H_0 : M_1 = M_2$

(2) Alternative Hypothesis

$H_1 : M_1 \neq M_2$ (two tail test)

(3) Level of significance

$\alpha = 0.05$

$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

degrees of freedom $= n_1 + n_2 - 2$

$= 30 - 2$

$= 28$

from table $t_{0.025}$ with 28 degrees of freedom is 2.048

(4) Test statistic $t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$= \frac{16(10)^2 + 14(8)^2}{28}$

$= 89.1$

$s = \sqrt{89.1} = 9.43$

(5) Conclusion $t_{cal} > t_{tab}$, reject H_0 and accept H_1

3) Majority specimens of nylon yarn taken from two machines, It was found that 8 specimens from first machine add a mean diner of 9.67 with a standard deviation of 1.81, while 10 specimens from second machine had a mean diner of 7.43 with a standard deviation of 1.48. Assuming that the proportions are normal, test the hypothesis $H_0 = M_1 - M_2 = 1.5$ against the alternative hypothesis $M_1 - M_2 > 1.5$ at 5% level of significance

Sol:- $n_1 = 8$ $n_2 = 10$

$\bar{x} = 9.67$ $s_1 = 1.81$

$\bar{y} = 7.43$ $s_2 = 1.48$

(1) Null Hypothesis

$H_0 : M_1 - M_2 = 1.5$

(2) Alternative Hypothesis

$H_1 : M_1 - M_2 > 1.5$

(right tailed test)

(3) $\alpha = 0.05$

degrees of freedom $= n_1 + n_2 - 2$

$= 16$

from table $t_{0.05}$ with 16 degrees of freedom is 1.746

(4) Test statistic

$t = \frac{\bar{x} - \bar{y} - (M_1 - M_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$= \frac{8(1.81)^2 + 10(1.48)^2}{16}$

$= 3.007$

$s = \sqrt{3.007} = 1.734$

$$t = \frac{(9.67 - 7.43) - (1.5)}{1.734\sqrt{\frac{1}{8} + \frac{1}{10}}}$$

$$= 0.8997$$

(5) Conclusion

$t_{cal} < t_{tab}$, accept H_0

$\therefore M_1 - M_2 = 1.5$

14) Find the maximum difference that we can expect with probability 0.95 between the means of samples of sizes 10 and 12 from a normal population if their standard deviations are found to be 2, 3 respectively.

Soln: Given that $n_1 = 10 \quad n_2 = 12$

$s_1 = 2 \quad s_2 = 3$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{10(4) + 12(9)}{20}$$

$$= \frac{148}{20} = 7.4$$

$$s = \sqrt{7.4} = 2.72$$

$\alpha = 0.05, \frac{\alpha}{2} = 0.025$

$t_{0.025}$ with 20 degrees of freedom is 2.086

The maximum difference between two means

$$E_{max} = t_{\alpha/2}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (2.086)(2.72)\sqrt{\frac{1}{10} + \frac{1}{12}}$$

$$= 2.429$$

15) The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered who have been drawn from the same normal population.

Soln: $n_1 = 9 \quad n_2 = 7$

$\bar{x} = 196.42$

$\bar{y} = 198.82$

$\sum_{i=1}^{9} (x_i - \bar{x})^2 = 26.94$

$\sum_{i=1}^{7} (y_j - \bar{y})^2 = 18.73$

$$s^2 = \frac{\sum_{i=1}^{9} (x_i - \bar{x})^2 + \sum_{j=1}^{7} (y_j - \bar{y})^2}{n_1 + n_2 - 2}$$

$$= \frac{26.94 + 18.73}{9 + 7 - 2}$$

$$= 3.26$$

$$s = \sqrt{3.26} = 1.81$$

(1) $H_0 : M_1 = M_2$

(2) $H_1 : M_1 \neq M_2$ (Two tail test)

(3) $\alpha = 0.05, \frac{\alpha}{2} = 0.025$

from table $t_{0.025}$ with 14 degrees of freedom is 2.145 ($\because n_1 + n_2 - 2 = 16 - 2 = 14$)

(4) Test statistic $t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$= \frac{196.42 - 198.82}{1.81\sqrt{\frac{1}{9} + \frac{1}{7}}}$$

$$= -2.63$$

(5) conclusion $|t_{cal}| > t_{tab}$

reject H_0, accept H_1

6) Two horses A and B were tested according to the time in seconds to run a particular track with the following results.

No. of trials (n)

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

Sol:

Horse A	Horse B	$x-\bar{x}$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
28	29	-3.286	10.8	0.84	0.70
30	30	-1.286	1.65	1.84	3.38
32	30	0.714	0.509	1.84	3.38
33	24	1.714	2.93	-4.16	17.3
33	27	1.714	2.93	-1.16	1.34
29	29	-2.286	5.22	0.84	0.70
34		2.714	7.36		
			31.4		26.8

$n_1 = 7$

$n_2 = 6$

$\bar{x} = \frac{28+30+32+33+33+29+34}{7}$

$\bar{x} = 31.286$

$\bar{y} = \frac{29+30+30+24+27+29}{6}$

$= 28.16$

$\sum_{i=1}^{7} (x_i - \bar{x})^2 = 31.4$

$\sum_{j=1}^{6} (y_j - \bar{y})^2 = 26.8$

$S^2 = \frac{\sum_{i=1}^{7} (x_i - \bar{x})^2 + \sum_{j=1}^{6} (y_j - \bar{y})^2}{n_1 + n_2 - 2}$

$= \frac{31.4 + 26.8}{11} = 5.29$

$S = \sqrt{5.29} = 2.3$

① $H_0 : M_1 = M_2$

② $H_1 : M_1 \neq M_2$ (Two tail test)

③ $\alpha = 0.05, \frac{\alpha}{2} = 0.025$

from table $t_{0.025}$ with $n_1 + n_2 - 2$ i.e., $7+6-2 = 13-2 = 11$ degrees of freedom is 2.20

④ Test statistic $t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$= \frac{31.286 - 28.16}{2.3\sqrt{\frac{1}{7} + \frac{1}{6}}}$

$= 2.443$

⑤ Conclusion

$t_{cal} > t_{tab}$

reject H_0 and accept H_1

Two horses does not have the same running capacit

7) To examine the hypothesis that the husbands are more intelligent than the wives. An investigator took a sample of 10 couples and administrated them a test which measures the IQ. The results are as follows.

Husbands	117	105	97	105	123	109	86	78	103	107
Wives	106	98	87	104	116	95	90	69	108	85

Test the hypothesis with a reasonable test at 5% Level of significance?

Sol:- $n_1 = 10 \quad n_2 = 10$

$$\bar{x} = \frac{117+105+97+105+123+109+86+78+103+107}{10}$$

$$\bar{x} = 103$$

$$\bar{y} = \frac{106+98+87+104+116+95+90+69+108+85}{10}$$

$$\bar{y} = 95.8$$

Husbands	Wives	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
117	106	14	196	10.2	104.04
105	98	2	4	2.2	4.84
97	87	-6	36	-8.8	77.44
105	104	2	4	8.2	67.24
123	116	20	400	20.2	408.04
109	95	6	36	-0.8	0.64
86	90	-17	289	-5.8	33.64
78	69	-25	625	-26.8	718.24
103	108	0	0	12.2	148.84
107	85	4	16	-10.8	116.64
			1606		1679.6

$$\sum_{i=1}^{10}(x_i-\bar{x})^2 = 1606$$

$$\sum_{i=1}^{10}(y_i-\bar{y})^2 = 1679.6$$

$$s^2 = \frac{\sum_{i=1}^{10}(x_i-\bar{x})^2 + \sum_{i=1}^{10}(y_i-\bar{y})^2}{n_1+n_2-2}$$

$$= \frac{1606+1679.6}{10+10-2}$$

$$= 182.53$$

$$S = \sqrt{182.53} = 13.51$$

(1) $H_0 : M_1 = M_2$

(2) $H_1 : M_1 > M_2$ (right tail test)

(3) $\alpha = 0.05$

from table $t_{0.05}$ with n_1+n_2-2 i.e., $10+10-2 = 18$ degrees of freedom is 1.734

(4) Test statistic $t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$$= \frac{103-95.8}{13.51\sqrt{\frac{1}{10}+\frac{1}{10}}}$$

$$= 1.19$$

(5) conclusion

$t_{cal} < t_{tab}$

accept H_0

∴ Husbands are more intelligent than wives.

F-Test

procedure

Test for equality of two population Variances

Let two independent random samples of sizes n_1 and n_2 be drawn from two normal populations.

To test the hypothesis that the population variances are equal or not (σ_1^2 and σ_2^2).

Let the Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$

Then the Alternative Hypothesis

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

The estimates of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1}$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{\sum_{i=1}^{n_2} (y_j - \bar{y})^2}{n_2 - 1}$$

where s_1^2 and s_2^2 are the variances of two samples.

The test statistic $F = \frac{S_1^2}{S_2^2}$ if

$S_1^2 > S_2^2$ $F_\alpha(\nu_1, \nu_2)$

$= F_\alpha(n_1 - 1, n_2 - 1)$

The test statistic $F = \frac{S_2^2}{S_1^2}$ if

$S_2^2 > S_1^2$ $F_\alpha(\nu_2, \nu_1)$

$= F_\alpha(n_2 - 1, n_1 - 1)$

if the calculated value of F > the tabulated value of F at 5% Level of significance.

∴, we reject H_0 (Null Hypothesis) and conclude that the variances σ_1^2 and σ_2^2 are not equal.

Otherwise we accept the Null Hypothesis H_0 and conclude that σ_1^2 and σ_2^2 are equal.

$F_\alpha(\nu_1, \nu_2)$ is the value of F with ν_1, ν_2 degrees of freedom such that the area under F distribution to the right of F_α is α.

In tables, F_α is tabulated for 5% and 1% Level of significance for various combination of the degrees of freedom ν_1 and ν_2. Clearly value of F at 5% Level of significance is lower than at 1% Level of significance (LOS).

1) In 1 sample of 8 observations from a normal population, the sum of the squares of the deviations of the sample values from the sample mean is 84.4 and in another sample of 10 observation it falls 102.6. Test at 5% level of significance whether the populations have the same variance.

Sol:- $n_1 = 8$ $\sum_{i=1}^{8} (x_i - \bar{x})^2 = 84.4$

$n_2 = 10$ $\sum_{j=1}^{10} (y_j - \bar{y})^2 = 102.6$

The estimators of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1}$$

$$= \frac{84.4}{7}$$

$$= 12.057$$

$$S_2^2 = \frac{\sum_{j=1}^{n_2} (y_j - \bar{y})^2}{n_2 - 1}$$

$$= \frac{102.6}{9}$$

$$= 11.4$$

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

(2) Alternative Hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

(3) $\alpha = 5\%$.

$F_{0.05}(n_1 - 1, n_2 - 1)$

$= F_{0.05}(7, 9) = 3.29$

(4) Test statistic

Since $S_1^2 > S_2^2$

$$F = \frac{S_1^2}{S_2^2} = \frac{12.057}{11.4} = 1.057$$

(5) conclusion

$F_{cal} < F_{tab}$

accept H_0, The populations have same variance.

***Note:-

F-test is for variance

When difference of two means of t-test should be applied if mean is mentioned in question.

2) In 1 Sample of 10 observations from a normal population the sum of the squares of the deviations of the sample values from the sample mean is 102.4 and in another sample of 12 observations from another normal population, the sum of squares of deviations of the sample values from the sample mean is 120.5. Examining whether the normal populations have the same variance.

Sol: $n_1 = 10$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 102.4$$

$n_2 = 12$

$$\sum_{i=1}^{12} (y_j - \bar{y})^2 = 120.5$$

The estimators of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1}$$

$$= \frac{102.4}{9} = 11.37$$

$$S_2^2 = \frac{\sum_{j=1}^{n_2} (y_j - \bar{y})^2}{n_2 - 1}$$

$$= \frac{120.5}{11} = 10.95$$

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

(2) Alternative Hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

(3) $\alpha = 5\%$.

$F_{0.05}(n_1 - 1, n_2 - 1)$

$= F_{0.05}(9, 11)$

$= 2.90$

(4) Test Statistic

Since $S_1^2 > S_2^2$

$$F = \frac{S_1^2}{S_2^2} = \frac{11.37}{10.95} = 1.038$$

(5) conclusion

$F_{cal} < F_{tab}$

accept H_0, The populations have same variance.

3) It is known that the mean diameters of rivets produced by 2 firms A and B are practically the same, but the standard deviations may differ. For 22 rivets produced by A, the standard deviation is 2.9 mm. While for 16 rivets manufactured by B, the standard deviation is 3.8 mm. Compute the statistic you could use to test whether the products of firm A have the same variability has those of firm B and test its significance.

Soln $n_1 = 22$ $n_2 = 16$

$s_1 = 2.9\text{mm}$ $s_2 = 3.8\text{mm}$

The estimators of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{22(2.9)^2}{22-1} = 8.805$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{16(3.8)^2}{16-1} = 15.393$$

(1) Null Hypothesis

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

(2) Alternative Hypothesis

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

(3) $\alpha = 5\%$ $F_{0.05}(n_2 - 1, n_1 - 1)$

$F_{0.05}(15, 21) = 2.18$

(4) Test statistic

since $S_1^2 < S_2^2$

$$F = \frac{S_2^2}{S_1^2} = \frac{15.393}{8.805} = 1.74$$

(5) Conclusion

$F_{cal} < F_{tab}$

accept H_0, The populations have same variance.

4) Pumpkins are grown under experimental conditions, two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal. Test Hypothesis that the two variances are equal

Soln $n_1 = 11$ $s_1 = 0.8$

$n_2 = 9$ $s_2 = 0.5$

The estimators of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{11(0.8)^2}{11-1} = 0.704$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{9(0.5)^2}{9-1} = 0.28$$

(1) Null Hypothesis

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

(2) Alternative Hypothesis

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

(3) $\alpha = 5\%$ $F_{0.05}(n_1 - 1, n_2 - 1)$

$F_{0.05}(10, 8) = 3.35$

(4) Test statistic

since $S_1^2 > S_2^2$

$$F = \frac{S_1^2}{S_2^2} = \frac{0.704}{0.28} = 2.51$$

(5) Conclusion

$F_{cal} < F_{tab}$

accept H_0, The populations have same variance

5) Time taken by the workers 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 from the populations from which this samples are drawn do not differ significantly.

Sol:- Given $n_1 = 6$ $n_2 = 7$

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6}$$

$$= \frac{20+16+26+27+23+22}{6}$$

$$\bar{x} = 22.3$$

$$\bar{y} = \frac{\sum_{j=1}^{7} y_j}{7}$$

$$= \frac{27+33+42+35+32+34+38}{7}$$

$$\bar{y} = 34.4$$

x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
20	−2.3	5.29	27	−7.4	54.76
16	−6.3	39.69	33	−1.4	1.96
26	3.7	13.69	42	7.6	57.76
27	4.7	22.09	35	0.6	0.36
23	0.7	0.49	32	−2.4	5.76
22	−0.3	0.09	34	−0.4	0.16
			38	3.6	12.96
		81.34			133.72

$$\sum_{i=1}^{6} (x_i - \bar{x})^2 = 81.34$$

$$\sum_{j=1}^{7} (y_j - \bar{y})^2 = 133.72$$

The estimators of σ_1^2 and σ_2^2 are given by

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1} = \frac{81.34}{5}$$

$$= 16.26$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2} (y_j - \bar{y})^2}{n_2 - 1} = \frac{133.72}{6}$$

$$= 22.29$$

(1) Null Hypothesis

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

(2) Alternative Hypothesis

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

(3) $\alpha = 5\%$

$F_{0.05}(n_2 - 1, n_1 - 1)$

$= F_{0.05}(6, 5)$

$= 4.95$

(4) Test statistic

since $s_1^2 < s_2^2$

$$F = \frac{s_2^2}{s_1^2} = \frac{22.89}{16.26}$$

$$= 1.37$$

(5) Conclusion

$F_{cal} < F_{tab}$

Accept H_0, The Population have same variance.

6) The measurements of the output of two units have given the following results. Assuming that both samples have been obtained from the normal population at 1% Level of significance, Test whether the two populations have the same variance.

Unit-A	14.1	10.1	14.7	13.7	14
Unit-B	14.	14.5	13.7	12.7	14.1

Sol: Given $n_1 = 5 \quad n_2 = 5$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5}$$

$$= \frac{14.1+10.1+14.7+13.7+14}{5}$$

$$= 13.32$$

$$\bar{y} = \frac{\sum_{j=1}^{5} y_j}{5}$$

$$= \frac{14+14.5+13.7+12.7+14.1}{5}$$

$$\bar{y} = 13.8$$

x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
14.1	0.78	0.608	14	0.2	0.04
10.1	−3.22	10.368	14.5	0.7	0.49
14.7	1.38	1.904	13.7	−0.1	0.01
13.7	0.38	0.1444	12.7	−1.1	1.21
14	0.68	0.462	14.1	0.3	0.09
		13.49			1.84

$$\sum_{i=1}^{5} (x_i - \bar{x})^2 = 13.49$$

$$\sum_{j=1}^{5} (y_j - \bar{y})^2 = 1.84$$

The estimators of σ_1^2 and σ_2^2 are given by

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1}$$

$$= \frac{13.49}{4}$$

$$= 3.372$$

$$S_2^2 = \frac{\sum_{j=1}^{n_2} (y_j - \bar{y})^2}{n_2 - 1}$$

$$= \frac{1.84}{4} = 0.46$$

(1) Null Hypothesis

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

(2) Alternative Hypothesis

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

(3) $\alpha = 5\%$

$F_{0.05}(n_1 - 1, n_2 - 1)$

$= F_{0.05}(4, 4)$

$= 6.39$

(4) Test statistic

Since $S_1^2 > S_2^2$

$$F = \frac{S_1^2}{S_2^2} = \frac{3.372}{0.46}$$

$$= 7.33$$

(5) Conclusion

$F_{cal} > F_{tab}$,

Reject H_0, The Populations does not have same variance.

χ^2 Test for independence of attributes

Literally, an attribute means a quality or characteristic. Examples of attributes are drinking, smoking, blindness, Honesty, beauty etc.

An attribute may be marked by its presence or absence in a number of given population.

Let the observations be classified according to two attributes and the frequencies O_i in the different categories be shown in a two way table which is known as contigency table.

We have to test on the basis of self frequencies whether the two attributes are independent or not.

We take the Null Hypothesis H_0 as there is no association between the attributes that is we assume that the two attributes are independent.

The Expected frequency (E_i)

$$= \frac{\text{row total} \times \text{column total}}{\text{Grand total}}$$

The test statistic $\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$

approximately follows χ^2 distribution with degrees of freedom (df)

$= (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)$

1) On the basis of information given below about the treatment of 200 patients suffering from a disease. State whether the new treatment is comparatively Superior to the conventional treatment.

	Favourable	Not Favourable	Total
New	60	30	90
Conventional	40	70	110
	100	100	200

Sol:- The expected frequencies (E_i) can be obtained from

$$\frac{\text{row total} \times \text{Column total}}{\text{Grand total}}$$

	Favourable	Not Favourable	Total
New	60	30	90
Conventional	40	70	110
	100	100	200

	Favourable	Not Favourable	
New	$\frac{90 \times 100}{200} = 45$	$\frac{90 \times 100}{200} = 45$	90
Conventional	$\frac{110 \times 100}{200} = 55$	$\frac{100 \times 110}{200} = 55$	110
	100	100	200

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
60	45	15	225	$\frac{225}{45} = 5$
30	45	−15	225	5
40	55	−15	225	4.09
70	55	15	225	4.09
		0		18.18

① Null Hypothesis: There is no difference between new and conventional treatment

② Alternative Hypothesis: New treatment is superior to the conventional treatment.

③ Level of significance $\alpha = 5\%$.

$d.f = (\text{no. of rows} - 1) \times (\text{no. of columns} - 1)$

$= (2-1) \times (2-1) = 1$

$\chi^2_{0.05}$ with 1 df is 3.841

④ Test statistic $\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$

$= 18.18$

⑤ $\chi^2_{cal} > \chi^2_{tab}$, reject H_0, accept H_1

New treatment is superior to the conventional treatment.

2) The following table gives the classification of 100 workers according to gender and nature of the work. Test whether the nature of the work independent of the gender.

	Stable	unstable	Total
Males	40	20	60
Females	10	30	40
	50	50	100

Sol:-

	Stable	unstable	Total
Males	$\frac{60 \times 50}{100} = 30$	$\frac{60 \times 50}{100} = 30$	60
Females	$\frac{40 \times 50}{100} = 20$	$\frac{40 \times 50}{100} = 20$	40
	50	50	

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
40	30	10	100	$\frac{100}{30} = 3.33$
20	30	−10	100	$\frac{100}{30} = 3.33$
10	20	−10	100	$\frac{100}{20} = 5$
30	20	10	100	$\frac{100}{20} = 5$
		0		16.66

① Null Hypothesis: The nature of the work is independent of gender.

② Alternative Hypothesis: The nature of the work is dependent of gender.

③ Level of significance

$\alpha = 5\%$

$df = (2-1)(2-1) = 1$

$\chi^2_{0.05}$ with 1df is 3.841

④ Test statistic $\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$

$= 16.66$

⑤ $\chi^2_{cal} > \chi^2_{tab}$, reject H_0, accept H_1

3) Given the following contigency table for Hair colour and Eye colour.

		Hair colour		
		Fair	Brown	Black
Eye colour	Blue	15	5	20
	Grey	20	10	20
	Brown	25	15	20

Find the value of χ^2?

Is there a good association between the two.

Sol:-

		Hair colour			Total
		Fair	Brown	Black	
Eye colour	Blue	15	5	20	40
	Grey	20	10	20	50
	Brown	25	15	20	60
		60	30	60	150

		Hair Colour			Total
		Fair	Brown	Black	
Eye colour	Blue	$\frac{40\times60}{150}=16$	$\frac{40\times30}{150}=8$	$\frac{60\times40}{150}=16$	40
	Grey	$\frac{50\times60}{150}=20$	$\frac{50\times30}{150}=10$	$\frac{50\times60}{150}=20$	50
	Brown	$\frac{60\times60}{150}=24$	$\frac{60\times30}{150}=12$	$\frac{60\times60}{150}=24$	60
		60	30	60	150

O_i	E_i	O_i-E_i	$(O_i-E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
15	16	−1	1	$\frac{1}{16}$
5	8	−3	9	$\frac{9}{8}$
20	16	4	16	$\frac{16}{16}$
20	20	0	0	0
10	10	0	0	0
20	20	0	0	0
25	24	1	1	$\frac{1}{24}$
15	12	3	9	$\frac{9}{12}$
20	24	−4	16	$\frac{16}{24}$
				3.645

① H_0 : There is no association between hair and eye colour

② H_1 : There is goal association between them

③ $\alpha = 5\%$.

$df = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)$

$= (3-1)(3-1) = 4$

$\chi^2_{0.05}$ with 4 df is 9.488

④ Test statistic $\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$

$= 3.645$

⑤ $\chi^2_{cal} < \chi^2_{tab}$, accept H_0

There is no association between hair and eye colour.

4) From the following data, find whether any significant liking in the habit of taking softdrinks among the categories of Employees. Use χ^2 distribution test with level of significance of 5%.

Soft drinks	Employees		
	Clerks	Teachers	officers
Pepsi	10	25	65
Thumsup	15	30	65
Fanta	50	60	30

soft

Soft drinks	Employees			
	Clerks	Teachers	officers	Total
Pepsi	10	25	65	100
Thumsup	15	30	65	110
Fanta	50	60	30	140
Total	75	115	160	350

Soft drinks	Employees			
	Clerks	Teachers	officers	Total
Pepsi	$\frac{100\times75}{350}=21.43$	$\frac{100\times115}{350}=32.9$	$\frac{100\times160}{350}=45.7$	100
Thumsup	$\frac{75\times110}{350}=23.6$	$\frac{115\times110}{350}=36.1$	$\frac{160\times110}{350}=50.3$	110
Fanta	$\frac{75\times140}{350}=30$	$\frac{115\times140}{350}=46$	$\frac{160\times140}{350}=64$	140
Total	75	115	160	350

① H_0 : There is no significant liking in the habit of taking softdrinks among the categories of Employees

(2) H_1: There is significant liking in the habit of taking softdrinks among the categories of Employees.

(3) $\alpha = 5\%$.

df = (No. of rows − 1) × (No. of columns − 1)

= (3−1)(3−1) = 4

$\chi^2_{0.05}$ with 4 df is 9.488

(4) Test statistic

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

= 60.25

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
10	21.43	−11.43	130.65	6.096
25	32.9	−7.9	62.41	1.9
65	45.7	19.3	372.49	8.15
15	23.6	−8.6	73.96	3.133
30	36.1	−6.1	37.21	1.03
65	50.3	14.7	216.09	4.3
50	30	20	400	13.33
60	46	14	196	4.26
30	64	−34	1156	18.06
				60.25

(5) $\chi^2_{cal} > \chi^2_{tab}$, accept H_1

There is significant liking in the habit of taking softdrinks among the categories of Employees.

Correlation and Regression

Previously, we are used to study the characteristics of only one variable marks, weights, prices, heights, ages, sales etc. This type of analysis uni-variate analysis. We may come accross certain series where each term of the series may assume values of two or more variables.

For example: If we measure the heights and weights of a certain group of people, we shall get what is known as bi-variate distribution.

In a bi-variate distribution, we may be interest to find out if there is any correlation or co-variance between the two variables under study.

If the change in 1 variable effects a change in the other variables, the variables are said to be correlated.

If the variables are deviated in the same direction that is if the increase or decrease of one results in the corresponding second results increase or decrease then we can say the correlation is direct or positive.

If they constantly deviated in the opposite direction, i.e., if increase in one variable corresponding decrease in other variable then the corelation is said to be diverse or negative.

For Example :-

The corelation between the heights and weights of a group of persons. The income and Expenditure are examples for positive correlation.

price and Demand of a commodity. The volume and pressure of a perfect gas are negatively corelated.

Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Karl Piarson coefficient of correlation

As a measure of intensity or degree of linear relationship between two variables, karl piarson - a british biometrisian developed a formula called correlation coefficient

Correlation coefficient between two variables X and Y is usually denoted by r and is defined as

$$r = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

Cov - covariance

$$Cov(X,Y) = E((x-\bar{x})(y-\bar{y}))$$

$$= \frac{\sum_{i=1}^{n} dx\,dy}{n}$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{n}$$

$$\sigma_x^2 = E((x-\bar{x})^2) = \frac{\sum dx^2}{n};$$

$$\sigma_x = \sqrt{\frac{\sum dx^2}{n}}$$

$$\sigma_y^2 = E((y-\bar{y})^2) = \frac{\sum dy^2}{n};$$

$$\sigma_y = \sqrt{\frac{\sum dy^2}{n}}$$

$$E(x) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Take $dx = x - \bar{x}$

$dy = y - \bar{y}$

$$\sigma_x \sigma_y = \sqrt{\frac{\sum dx^2}{n}\frac{\sum dy^2}{n}} = \sqrt{\frac{\sum dx^2 \sum dy^2}{n^2}}$$

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\frac{\sum dxdy}{n}}{\sqrt{\frac{\sum dx^2 \sum dy^2}{n^2}}}$$

$$= \frac{\sum dxdy}{\sqrt{\sum dx^2 \sum dy^2}}$$

Properties of correlation coefficient

- The coefficient of correlation lies between -1 and $+1$ i.e., $-1 \le r \le 1$ (or) $|r| \le 1$
- If $r = 1$, The correlation is perfect and positive.
- If $r = -1$ The correlation is perfect and negative
- If $r = 0$, then there is no relationship between two variables i.e., the two variables are independent.
- The coefficient of correlation is independent of the change of origin and scale of measurements.

1) Calculate the coefficient of correlation from the following data.

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Soln:

x	y	$dx = x-\bar{x}$	$dy = y-\bar{y}$	dxdy	dx^2	dy^2
12	14	2	5	10	4	25
9	8	−1	−1	1	1	1
8	6	−2	−3	6	4	9
10	9	0	0	0	0	0
11	11	1	2	2	1	4
13	12	3	3	9	9	9
7	3	−3	−6	18	9	36
		0	0	46	28	84

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

$$\bar{y} = \frac{\sum_{i=1}^{7} y_i}{7} = \frac{63}{7} = 9$$

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 \sum dy^2}} = \frac{46}{\sqrt{28 \times 64}} = 0.94$$

2) Find if there is any significant correlation between the Heights and weights given below.

Heights (in inches)	57	59	62	63	64	65	55	58	57
Weights (in lbs)	113	117	126	126	130	129	111	116	112

Soln:- Heights − x
Weights − y

x	y	$dx = x-\bar{x}$	$dy = y-\bar{y}$	dxdy	dx^2	dy^2
57	113	−3	−7	21	9	49
59	117	−1	−3	3	1	9
62	126	2	6	12	4	36
63	126	3	6	18	9	36
64	130	4	10	40	16	100
65	129	5	9	45	25	81
55	111	−5	−9	45	25	81
58	116	−2	−4	8	4	16
57	112	−3	−8	24	9	64
		0	0	216	102	472

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i}{9} = 60$$

$$\bar{y} = \frac{\sum_{i=1}^{9} y_i}{9} = \frac{1080}{9} = 120$$

$$r = \frac{\sum dx\,dy}{\sqrt{\sum dx^2 \sum dy^2}} = \frac{216}{\sqrt{102 \times 472}} = 0.984$$

3) Find Karl - Piarson coefficient of correlation from the following data.

wages	100	101	102	102	100	99	97	98	96	95
cost of living	98	99	99	97	95	92	95	94	90	91

Sol:- wages $-x$
cost of living $-y$

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dxdy	dx^2	dy^2
100	98	1	3	3	1	9
101	99	2	4	8	4	16
102	99	3	4	12	9	16
102	97	3	2	6	9	4
100	95	1	0	0	1	0
99	92	0	-3	0	0	9
97	95	-2	0	0	4	0
98	94	-1	-1	1	1	1
96	90	-3	-5	15	9	25
95	91	-4	-4	16	16	16
		0	0	61	54	96

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 99$$

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 95$$

$$r = \frac{\sum dx\,dy}{\sqrt{\sum dx^2 \sum dy^2}} = \frac{61}{\sqrt{54 \times 96}} = 0.847$$

4) Calculate the coefficient of correlation between ages of cars and annual maintenance cost.

age of cars (in years)	2	4	6	7	8	10	12
annual maintenance (in rupees)	1600	1500	1800	1900	1700	2100	2000

Sol:- age of cars $-x$
annual maintenance $-y$

$$\bar{x} = \frac{49}{7} = 7$$

$$\bar{y} = \frac{12600}{7} = 1800$$

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dxdy	dx^2	dy^2
2	1600	-5	-200	1000	25	40000
4	1500	-3	-300	900	9	90000
6	1800	-1	0	0	1	0
7	1900	0	100	0	0	10000
8	1700	1	-100	-100	1	10000
10	2100	3	300	900	9	90000
12	2000	5	200	1000	25	40000
		0	0	3700	70	280000

$$r = \frac{3700}{\sqrt{70 \times 280000}} = 0.835$$

Rank correlation coefficient

This method is based on ranks and is useful in dealing with qualitative characteristics such as morality, character, intelligence and beauty. It cannot be measured quantitatively as in the case of piarson coefficient correlation.

Rank correlation is applicable only to the individual observations.

The formula for spearman's rank correlation is

$$\rho = 1 - \left\{ \frac{6\Sigma D^2}{N(N^2-1)} \right\}$$

1) 10 competitors in a musical test were ranked by 3 judges A, B and C in the following order.

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach, two common likings in music.

R_A	R_B	R_C	$D_1 = R_A - R_B$	$D_2 = R_A - R_C$	$D_3 = R_B - R_C$	D_1^2	D_2^2	D_3^2
1	3	6	−2	−5	−3	4	25	9
6	5	4	−1	2	1	1	4	1
5	8	9	−3	−4	−1	9	16	1
10	4	8	6	2	−4	36	4	16
3	7	1	−4	2	6	16	4	36
2	10	2	−8	0	8	64	0	64
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	−1	1	2	1	1	4
						200	60	214

$$\rho(A,B) = 1 - \frac{6\Sigma D_1^2}{N(N^2-1)}$$

$$= 1 - \left\{ \frac{6 \times 200}{10(100-1)} \right\}$$

$$= -0.21$$

$$\rho(A,C) = 1 - \frac{6\Sigma D_2^2}{N(N^2-1)}$$

$$1 - \left\{ \frac{6 \times 60}{10 \times 99} \right\}$$

$$= 0.63$$

$$\rho(B,C) = 1 - \frac{6\Sigma D_3^2}{N(N^2-1)}$$

$$= 1 - \left\{ \frac{6 \times 214}{10 \times 99} \right\}$$

$$= -0.29$$

Since rank of A, C is maximum we conclude that the pair of judges A, C has the nearest approach to common likings in music.

2) Following are the ranks obtained by 10 students in two different subjects statistics and mathematics. To what extinct the knowledge of students in two subjects is related.

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Sol:-

Statistics (Rs)	Mathematics (Rm)	D = Statistics − Mathematics (D = Rs − Rm)	D^2
1	2	−1	1
2	4	−2	4
3	1	+2	4
4	5	−1	1
5	3	+2	4
6	9	−3	9
7	7	0	0
8	10	−2	4
9	6	3	9
10	8	2	4
		0	40

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6(40)}{10(100-1)}$$

$$= 0.757$$

3) The ranks of 16 students in maths and statistics are given below.

Maths	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
statistics	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Calculate the rank correlation coefficient for proficiencies of this group in maths and statistics.