1.2.b t-distribution for difference of means

Suppose, we want to test, if two independent samples $x_1, x_2, \ldots x_{n_1}$, and $y_1, y_2, \ldots y_{n_2}$ of sizes n_1 and n_2 have been drawn from two normal population with means μ_1 and μ_2 respectively.

Under the null hypothesis, H_0 that the samples have been drawn from the normal population with means μ_1 and μ_2 are under the assumption that the population variance are equal. (i.e., $\sigma_1 = \sigma_2 = \sigma$)

The statistic t, where $t = \dfrac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

where, $\bar{x} = \dfrac{\Sigma x_i}{n_1}, \bar{y} = \dfrac{\Sigma y_i}{n_2}$

and $S^2 = \dfrac{1}{n_1 + n_2 - 2}[\Sigma (x_i - \bar{x})^2 + \Sigma (y_i - \bar{y})^2]$

is an unbiased estimate of the population variance σ^2.

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

Under the null hypothesis H_0, that (i) samples have been drawn from the population with the same means i.e., $\mu_1 = \mu_2$, or (ii) the sample means $\bar{x}$ and $\bar{y}$ do not differ significantly, take the statistic

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ with degrees of freedom, $n_1 + n_2 - 2$.

If $|t| <$ table value of t_α accept H_0, at α level of significance.

Assumption : The following assumptions are made in using this test.

(i) Parent populations, from which the samples have been drawn are normally distributed.

(ii) Population variances are equal and unknown

(iii) The two samples are random and independent.

Since, $$S^2 = \frac{[\Sigma (x - \bar{x})^2 + \Sigma (y - \bar{y})^2]}{n_1 + n_2 - 2}$$

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

s_1, s_2 are standard deviations of the two samples. Therefore, statistic 't' to be tested is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Note 1. If $n_1 = n_2 = n$, $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n - 1}}}$ can be used as a test statistic

Note 2. If the pairs of values are in some way associated (or correlated) we cannot adopt the case under Note 1. Then, we have to find the differences of the associated pairs of values and apply for single mean

i.e., $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}}$, to test, if the means of the differences is significantly different from zero. Then, the degree of freedom is $n - 1$.

Here, the test statistic is $t = \dfrac{\bar{d}}{s/\sqrt{n-1}}$ or $\dfrac{\bar{d}}{S/\sqrt{n}}$

where $\bar{d}$ is the mean of the paired differences.

i.e., $d_i = x_i - y_i$ and $\bar{d} = \bar{x} - \bar{y}$,

where (x_i, y_i) are the paired data, $i = 1, 2, \ldots n$

Working Procedure :

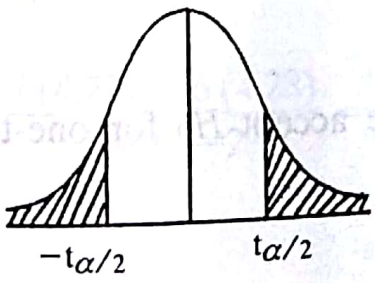
Concerning difference between two means, with unknown σ_1^2 and σ_2^2 but equal. $(\sigma_1 = \sigma_2 = \sigma)$. For the small samples $(n_1 < 30, n_2 < 30)$ drawn from two normal population.

1. Null hypothesis $H_0 : \mu_1 = \mu_2$
2. Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (or) $\mu_1 > \mu_2$

   (or) $\mu_1 < \mu_2$
3. Level of significance : α, d.f. $= n_1 + n_2 - 2$
4. Critical region :

(a) If $\mu_1 \neq \mu_2$, then the test is **two-tailed test** for the given α.

The critical values are $-t_{\alpha/2}$ and $t_{\alpha/2}$ from the t-distribution table with d.f. $= n_1 + n_2 - 2$

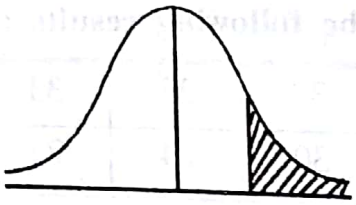


(b) If $\mu_1 > \mu_2$, then the test is **one-tailed test** (right) for the given α,

The critical value is t_α with d.f. $= n_1 + n_2 - 2$

3. $\alpha = 5\%$, d.f $= n_1 + n_2 - 2 = 7 + 6 - 2 = 11$

4. Critical region :



$t_\alpha = 1.796$

5. The test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{31.29 - 28.17}{\sqrt{(5.03)\left(\frac{1}{7} + \frac{1}{6}\right)}} = 2.498$$

6. Conclusion :

If $t < t_\alpha$, then we accept H_0 ; otherwise, we reject H_0.

Here, $2.498 \not< 1.796$

So, we reject H_0 at 5% level of significance.

Example 1.2b(2)

A group of 10 rats fed on diet A and another group of 8 rats fed on diet B, recorded the following increase in weight (gms).
Diet A : 5, 6, 8, 1, 12, 4, 3, 9, 6, 10
Diet B : 2, 3, 6, 8, 10, 1, 2, 8
Does it show superiority of diet A over diet B. **[A.U. N/D 2011]**

Solution : **Given :** $n_1 = 10, n_2 = 8$

$\Sigma x_1 = 5 + 6 + 8 + 1 + 12 + 4 + 3 + 9 + 6 + 10 = 64$

$\Sigma x_1^2 = 5^2 + 6^2 + 8^2 + 1^2 + 12^2 + 4^2 + 3^2 + 9^2 + 6^2 + 10^2 = 512$

$\Sigma x_2 = 2 + 3 + 6 + 8 + 10 + 1 + 2 + 8 = 40$

$\Sigma x_2^2 = 2^2 + 3^2 + 6^2 + 8^2 + 10^2 + 1^2 + 2^2 + 8^2 = 282$

$$\bar{x}_1 = \frac{\Sigma x_1}{10} = \frac{64}{10} = 6.4$$

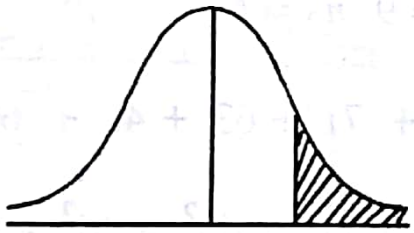
$$\bar{x}_2 = \frac{\Sigma x_2}{8} = \frac{40}{8} = 5$$

$$s_1^2 = \frac{\Sigma x_1^2}{n_1} - (\bar{x}_1)^2 = \frac{512}{10} - (6.4)^2 = 10.24$$

$$s_2^2 = \frac{\Sigma x_2^2}{n_2} - (\bar{x}_2)^2 = \frac{282}{8} - 25 = 10.25$$

$$S^2 = \left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right) = \frac{10\,(10.24) + 8\,(10.25)}{10 + 8 - 2} = 11.525$$

1. $H_0 : \mu_1 = \mu_2$
2. $H_1 : \mu_1 > \mu_2$ [One-tailed test (right)]
3. $\alpha = 5\%$ d.f. $= n_1 + n_2 - 2 = 10 + 8 - 2 = 16$
4. Critical region



5. The test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{6.4 - 5}{\sqrt{11.525\left(\frac{1}{10} + \frac{1}{8}\right)}}$$

$$= \frac{1.4}{1.6103} = 0.869$$

6. Conclusion :

If $t < t_\alpha$, then we accept H_0; otherwise, we reject H_0.

Here, $0.869 < 1.75$

So, we accept H_0.

Hence, the difference is not significant, so we cannot conclude that diet A is superior to diet B.

Example 1.2b(3)

The following are the number of sales which a sample of 9 salespeople of industrial chemicals in Gujarat and a sample of 6 sales people of industrial chemicals in Maharashtra made over a certain fixed period of time :

Gujarat :	59	68	44	71	63	46	69	54	48
Maharashtra :	50	36	62	52	70	41			

Assuming that the population sampled can be approximated closely with normal distributions having the same variance, test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ at the 0.01 level of signifance. **[A.U N/D 2009]**

Solution : Given : $n_1 = 9,\ n_2 = 6$

$$\Sigma x_1 = 59 + 68 + 44 + 71 + 63 + 46 + 69 + 54 + 48 = 522$$

$$\Sigma x_1^2 = 59^2 + 68^2 + 44^2 + 71^2 + 63^2 + 46^2 + 69^2 + 54^2 + 48^2 = 31148$$

$$\Sigma x_2 = 50 + 36 + 62 + 52 + 70 + 41 = 311$$

$$\Sigma x_2^2 = 50^2 + 36^2 + 62^2 + 52^2 + 70^2 + 41^2 = 16925$$

$$\bar{x}_1 = \frac{\Sigma x_1}{9} = \frac{522}{9} = 58$$

$$\bar{x}_2 = \frac{\Sigma x_2}{6} = \frac{311}{6} = 51.83$$

$$s_1^2 = \frac{\Sigma x_1^2}{n_1} - (\bar{x}_1)^2 = \frac{31148}{9} - (58)^2 = 96.89$$

$$s_2^2 = \frac{\Sigma x_2^2}{n_2} - (\bar{x}_2)^2 = \frac{16925}{6} - (51.83)^2 = 134.48$$

$$S^2 = \left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right] = \frac{9\,(96.89) + 6\,(134.48)}{9 + 6 - 2}$$

$$= \frac{872.01 + 806.88}{13} = 129.15$$

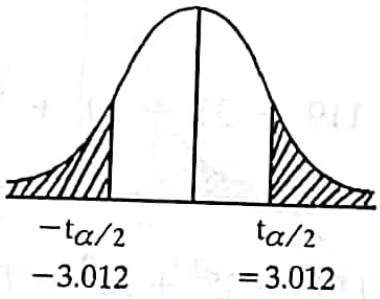
$$S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = (129.15)\left(\frac{1}{9} + \frac{1}{6}\right) = (129.15)\left(\frac{5}{18}\right) = 35.88$$

1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$ [Two-tailed test]

3. $\alpha = 1\%$, d.f. $= n_1 + n_2 - 2 = 9 + 6 - 2 = 13$

4. Critical region :



5. The test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{58 - 51.83}{\sqrt{35.88}} = 1.03$$

6. Conclusion :

If $-t_{\alpha/2} < t < t_{\alpha/2}$, then we accept H_0 ; otherwise, we reject H_0.

Here, $-3.102 < 1.03 < 3.012$

So, we accept H_0.

Example 1.2b(4)

The following are the average weekly losses of working hours due to accidents in 10 industrial plants before and after an introduction of a safety program was put into operation.

Before :	45	73	46	124	33	57	83	34	26	17
After :	36	60	44	119	35	51	77	29	24	11

Use to 0.05 level of significance to test whether the safety program is effective. **[A.U. N/D 2008]**

Solution :

$$\Sigma x_1 = 45 + 73 + 46 + 124 + 33 + 57 + 83 + 34 + 26 + 17$$

$$= 538$$

$$\Sigma x_1^2 = 45^2 + 73^2 + 46^2 + 124^2 + 33^2 + 57^2 + 83^2 + 34^2 + 26^2 + 17^2$$

$$= 38194$$

$$\Sigma x_2 = 36 + 60 + 44 + 119 + 35 + 51 + 77 + 29 + 24 + 11$$

$$= 486$$

$$\Sigma x_2^2 = 36^2 + 60^2 + 44^2 + 119^2 + 35^2 + 51^2 + 77^2 + 29^2 + 24^2 + 11^2$$

$$= 32286$$

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{538}{10} = 53.8$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{486}{10} = 48.6$$

$$s_1^2 = \frac{\Sigma x_1^2}{n_1} - (\bar{x}_1)^2 = \frac{38194}{10} - (53.8)^2$$

$$= 3819.4 - 2894.44$$

$$= 924.96$$

$$s_2^2 = \frac{\Sigma x_2^2}{n_2} - (\bar{x}_2)^2 = \frac{32286}{10} - (48.6)^2$$

$$= 3228.6 - 2361.96$$

$$= 866.64$$

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{(10)\,(924.96) + (10)\,(866.64)}{10 + 10 - 2}$$

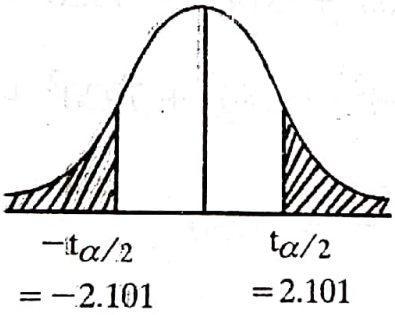
$$= \frac{9249.6 + 8666.4}{18} = 995.33$$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ [Two-tailed test]

$\alpha = 5\%$, d.f. $= n_1 + n_2 - 2 = 10 + 10 - 2 = 18$

Critical region :



The test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{53.8 - 48.6}{\sqrt{995.33\left(\frac{1}{10} + \frac{1}{10}\right)}} = 0.369$$

Conclusion :

If $-t_{\alpha/2} < t < t_{\alpha/2}$, then we accept H_0 ; otherwise, we reject H_0

Here, $-2.101 < 0.369 < 2.101$

So, we accept H_0.

The following random samples are measurements of the heat producing capacity (in millions of calories per ton) of specimen's of coals from two mines.

Mine 1:	8,260	8,130	8,350	8,070	8,340	
Mine 2:	7,950	7,890	7,900	8,140	7,920	7,840

Use the 0.01 level of significance to test whether the difference between the means of these two samples is significant.

[A.U. N/D 2008

Solution :

$\Sigma x_1 = 8260 + 8130 + 8350 + 8070 + 8340 = 41150$

$\Sigma x_1^2 = 8260^2 + 8130^2 + 8350^2 + 8070^2 + 8340^2 = 338,727,500$

$\Sigma x_2 = 7950 + 7890 + 7900 + 8140 + 7920 + 7840 = 47650$

$\Sigma x_2^2 = 7950^2 + 7890^2 + 7900^2 + 8140^2 + 7920^2 + 7840^2 = 378,316,200$

$$\bar{x}_1 = \frac{\Sigma x_1}{5} = \frac{41150}{5} = 8230$$

$$\bar{x}_2 = \frac{47640}{6} = 7940$$

$$s_1^2 = \frac{\Sigma x_1^2}{n_1} - (\bar{x}_1)^2 = \frac{338,727,500}{5} - (8230)^2$$

$$= 67745500 - 67732900$$

$$= 12,600$$

$$s_2^2 = \frac{\Sigma x_2^2}{n_2} - (\bar{x}_2)^2 = \frac{378316200}{6} - (7940)^2$$

$$= 63052700 - 63043600$$

$$= 9,100$$

$$S^2 = \left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right]$$

$$= \frac{(5)\,(12600) + (6)\,(9100)}{5 + 6 - 2}$$

$$= \frac{117600}{9}$$

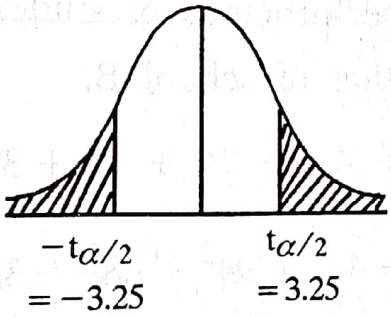
$$S^2 = 13066.67$$

1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$ [Two-tailed test]

3. $\alpha = 1\%$, d.f. $= n_1 + n_2 - 2 = 5 + 6 - 2 = 9$

4. Critical region :



5. The test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{8230 - 7940}{\sqrt{13066.67\left(\frac{1}{5} + \frac{1}{6}\right)}} = 4.19$$

6. Conclusion :

If $-t_{\alpha/2} < t < t_{\alpha/2}$, then we accept H_0 ; otherwise, we reject H_0.

Here, $-3.25 < 4.19 \not< 3.25$

So, we reject H_0.

1.3 TEST BASED ON χ^2-DISTRIBUTION

1.3. (a) χ^2-test for population variance

Let $x_1, x_2, ..., x_n$ be a random sample from a normal population with variance σ^2. Set the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$. Then the test statistic is $\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_0} \right)^2 = \frac{ns^2}{\sigma_0^2}$ where s^2 is the variance of the sample. Then $\chi^2 = \frac{ns^2}{\sigma_0^2}$ defined above follows a χ^2-distribution with $n - 1$ degrees of freedom.

By comparing the calculated value of χ^2 with the table value of χ^2 for $n - 1$ degrees of freedom at any required level of significance we may accept or reject the null hypothesis.

Note. If the sample size n is large ($n > 30$) then we can apply Fisher's approximation $\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1)$.

$\therefore Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0, 1)$ and we can apply normal test.

Example 1.3.a(1)

A random sample of size 25 from a population gives the sample standard deviation 8.5. Test the hypothesis that the population s.d is 10.

Solution : Given : $n = 25$, $s = 8.5$, $\sigma = 10$

1. $H_0 : \sigma = 10$
2. $H_1 : \sigma \neq 10$
3. $\alpha = 0.5\%$, d.f. $= n - 1 = 25 - 1 = 24$
4. Table value of $\chi^2 = 36.415$
5. Test test statistic

$$\chi^2 = \frac{n s^2}{\sigma^2} = \frac{(25)(8.5)^2}{(10)^2} = 18.06$$

6. **Conclusion :**

If calculated χ^2 < table χ^2, then we accept H_0 ; otherwise, we reject H_0.

Here, 18.06 < 36.415. So, we accept H_0.

Example 1.3.a.(2)

The s.d of the distribution of times taken by 15 workers for performing a job is 6.4 sec. Can it be taken as a sample from a population whose s.d is 5 sec ?

Solution : Given : $n = 15$, $s = 6.4$, $\sigma = 5$

1. $H_0 : \sigma = 5$
2. $H_1 : \sigma \neq 5$
3. $\alpha = 0.5\%$, d.f. $= n - 1 = 15 - 1 = 14$
4. Table value of $\chi^2 = 23.685$
5. Test statistic

$$\chi^2 = \frac{n s^2}{\sigma^2} = \frac{(15)\,(6.4)^2}{(5)^2} = 24.58$$

6. Conclusion :

If calculated χ^2 < table χ^2, then we accept H_0; otherwise, we reject H_0.

Here, 24.58 $\nless$ 23.685. i.e., 24.58 > 23.685. So, we reject H_0.

Example 1.3.a(3)

It is believed that the precision (as measured by the variance) of an instrument is no more than 0.16. Write down the null and alternative hypothesis for testing this belief. Carry out the test at 1% level given 11 measurements of the same subject on the instrument.

2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5

Solution : Given : $\sigma^2 = 0.16$

X	$X - \bar{X}$	$(X - \bar{X})^2$
2.5	–0.01	0.0001
2.3	–0.21	0.0441
2.4	–0.11	0.0121
2.3	–0.21	0.0441
2.5	–0.01	0.0001
2.7	+0.19	0.0361
2.5	–0.01	0.0001
2.6	+0.09	0.0081
2.6	+0.09	0.0081
2.7	+0.19	0.0361
2.5	–0.01	0.0001
$\bar{X} = \frac{27.6}{11} = 2.51$		$\Sigma (X - \bar{X})^2 = 0.1891$

1. $H_0 : \sigma^2 = 0.16$
2. $H_1 : \sigma^2 \neq 0.16$
3. $\alpha = 0.01$, d.f $= n - 1 = 11 - 1 = 10$
4. Table value $\chi^2 = 23.2$
5. **The test statistic is** $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma (X - \bar{X})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182$
6. **Conclusion :**

   If calculated, $\chi^2 <$ table χ^2, then we accept H_0 ; otherwise, we reject H_0.

Here, $\chi^2 = 1.182 < 23.2$, we accept H_0 at 1% level of significance.

We conclude that the data are consistent with the hypothesis that the precision of the instrument is 0.16.

Example 1.3.a.(4)

Test the hypothesis that $\sigma = 10$, given that s = 15 for a random sample of size 50 from a normal population.

Solution : Given : $\sigma = 10, s = 15, n = 50$

1. $H_0 : \sigma = 10$
2. $H_1 : \sigma \neq 10$
3. $\alpha =$ not given, choose $\alpha = 0.05$ and $\alpha = 0.01$
4. **The test statistic is** $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{(50)\,(15)^2}{100} = 112.5$

   Since n is large, use the test statistic in

   $$Z = \sqrt{2\chi^2} - \sqrt{2n-1} = \sqrt{225} - \sqrt{99} = 15 - 9.95 = 5.05$$

5. **Reject H_0 if** $|z| > 3$ for both the level of significance.
6. **Conclusion :**

Since, $|z| > 3$, so we reject H_0, it is significant at all levels of significance. We conclude that $\sigma \neq 10$

EXERCISES

1. A sample of 20 observations gave a standard deviation 3.72. Is this compatible with the hypothesis that the sample is from a normal population with variance 4.35.
2. A random sample of size 20 from a population gives the sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9.
3. Given 11 measurements of an instrument as 2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.6, 2.6, 2.7, 2.5. It is believed that the precision of that instrument as measured by the variance is 0.16. Test whether the data are consistent with the hypothesis (at 1% level of significance).

4. A sample of 12 values shows the s.d. to be 11. Does this agree with the hypothesis that the population s.d is 10, the population being normal ?

5. Weights in kgs of 10 students are given as 28, 40, 45, 53, 47, 43, 55, 48, 45, 49. Can we say that variance of the distribution of weights of all students from which the above sample of 10 students was drawn is equal to 20 sq.kgs.

1.3.b. χ^2-test to test the goodness of fit.

A very powerful test for testing the significance of the discrepancy between theory and experiment was given by prof. Karl-Pearson in 1990 and is known as "Chi-square test of goodness of fit". It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

■ I. Chi-Square Test for Goodness of fit ■

χ^2-test of goodness of fit is a test to find if the deviation of the experiment from theory is just by chance or it is due to the inadequacy of the theory to fit the observed data.

By this test, we test whether differences between observed and expected frequencies are significant or not.

χ^2-test statistic of goodness of fit is defined by

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}, \text{ where } \quad O \rightarrow \text{Observed frequency}$$

$$E \rightarrow \text{Expected frequency}$$

■ II. Application or uses of χ^2-distribution ■

(1) To test the "goodness of fit".

(2) To test the "independence of attributes".

(3) To test if the hypothetical value of the population variance is σ

(4) To test the homogeneity of independent estimates of the population variance.

(5) To test the homogeneity of independent estimates of the population correlation coefficient.

■ III. Conditions for the application of χ^2-test. ■

[AU N/D 2011]

(1) The sample observations should be independent.

(2) Constraints on the cell frequencies, if any, must be linear [e.g., $\Sigma O_i = \Sigma E_i$]

(3) N, the total frequency, should be atleast 50.

(4) No theoretical cell frequency should be less than 5.

■ IV. Independence of attributes ■

Note 1 In the case of

fitting a Binomial distribution, d.f $= n - 1$

fitting a Poisson distribution, d.f $= n - 2$

fitting a Normal distribution, d.f $= n - 3$

Note 2 If $\chi^2 = 0$, all observed and expected frequencies coincide

Note 3 For χ^2-distribution, mean $= \nu$, Variance $= 2\nu$

Example 1.3.b.(1)

Five coins are tossed 256 times. The number of heads observed is given below. Examine if the coins are unbiased, by employing χ^2 goodness of fit.

No. of heads	0	1	2	3	4	5
Frequency	5	35	75	84	45	12

Solution :

Given : $n = 6$, N = total number of frequencies $= 256$

1. H_0 : Binomial is a good fit.

2. H_1 : Binomial is not a good fit.

3. $\alpha = 0.05$, d.f $= n - 1 = 6 - 1 = 5$

4. Table value of $\chi^2 = 11.07$

5. **The test statistic is** $\chi^2 = \Sigma \frac{(O - E)^2}{E}$

On the assumption H_0, the expected frequencies are given by the terms of N $(q + p)^n = 256 \left(\frac{1}{2} + \frac{1}{2}\right)^5$

$$= \frac{256}{32}[5c_0 + 5c_1 + 5c_2 + 5c_3 + 5c_4 + 5c_5]$$

$$= \frac{256}{32}[1 + 5 + 10 + 10 + 5 + 1]$$

$$= 8\,[1 + 5 + 10 + 10 + 5 + 1]$$

$\therefore$ The expected frequencies are

8, 40, 80, 80, 40, 8

No. of heads	O	E	O – E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	5	8	–3	9	1.2500
1	35	40	–5	25	0.6250
2	75	80	–5	25	0.3125
3	84	80	4	16	0.2000
4	45	40	5	25	0.6250
5	12	8	4	16	2.000
	256	256			4.8875

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} = 4.8875$$

6. **Conclusion :**

If cal. χ^2 < table χ^2, then we accept H_0 ; otherwise, we reject H_0.

Here, $\chi^2 = 4.8875 < 11.07$. So, we accept to at 5% level of signiciance.

$\therefore$ Binomial distribution is a good fit to the given data.

Example 1.3.b.(2)

4 coins were tossed 160 times and the following results were obtained:

No. of heads :	0	1	2	3	4
Observed frequencies :	17	52	54	31	6

Under the assumption that the coins are unbiased, find the expected frequencies of getting 0, 1, 2, 3, 4 heads and test the goodness of fit.

Solution : **[AU A/M 2011]**

1. **Null hypothesis H_0** : The coins are unbiased
2. **Alternative hypothesis H_1** : The coins are biased
3. **Level of significance** : $\alpha = 0.05$, d.f. $= n - 1 = 5 - 1 = 4$
4. **Table value of χ^2** : 9.488
5. **Test statistic** : Under H_0, the test statistic is

$$\chi^2 = \sum \frac{(0-E)^2}{E} \sim \chi^2 \text{ distribution with } n-1 \text{ d.f.}$$

Probability of getting head $= p = \frac{1}{2}$

Probability of getting tail $= q = \frac{1}{2}$

Then the expected frequencies are

$$p(x) = n C_x p^x q^{n-x},\ x = 0, 1, 2, 3, \ldots$$

$$p(0 \text{ head}) = 4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = 0.0625$$

$$p(1 \text{ head}) = 0.25, \quad p(2 \text{ heads}) = 0.375$$

$$p(3 \text{ head}) = 0.25, \quad p(4 \text{ heads}) = 0.0625$$

χ^2 value is calculated from the following table :

No. of heads (x_i)	O	$p(x_i)$	$E = (160 \times p(x_i))$	$\frac{(O-E)^2}{E}$
0	17	0.0625	10	4.9
1	52	0.25	40	3.6
2	54	0.375	60	0.6
3	31	0.25	40	2.025
4	6	0.0625	10	1.6
Total	160		160	12.725

Calculated $\chi^2 = 12.725$

6. Conclusion :

If cal. $\chi^2 <$ table χ^2, then we accept H_0 ; otherwise, we reject H_0

Here, $12.725 \not< 9.488$

So, we reject H_0 and accept H_1

i.e., the coins are biased.

Example 1.3.b.(3)

A company keeps records of accidents. During a recent safety review, a random sample of 60 accidents was selected and classified by the day of the week on which they occured.

Day :	Mon	Tue	Wed	Thu	Fri
No. of accidents :	8	12	9	14	17

Test whether there is any evidence that accidents are more likely on some days than others.

Soluti:

	1	2	3	4	5	6	7	8	9	10
Price (in Rs.) :	115	118	120	140	135	137	139	142	144	150

Test the hypothesis that the expected price in the i^{th} month is Rs. $(100 + 3i)$, $i = 1, 2, \ldots 10$ with a standard deviation of Rs. 5 under the assumption that the prices are normally distributed.

8. To test a hypothesis H_0, an experiment is performed 3 times. The resulting values of chi-square are 2.37, 1.86 and 3.54, each of which corresponds to one degree of freedom. Show that while H_0 cannot be rejected at 5% level on the basis of any individual experiment, it can be rejected when the three experiments are collectively counted.

[Hint : Use additive property of chi-square variates]

1.3.c. χ^2-test to test the independence of attributes

Let us consider two attributes A and B. A divided in r classes $A_1, A_2, \ldots, A_r$ and B divided into S classes $B_1, B_2, \ldots, B_s$. If this is expressed as $r \times s$ matrix, the matrix is called $r \times s$ co-tingency table.

If (A_i), (B_j) represent the number of persons possessing the attribute A_i and B_j respectively $(i = 1, 2, \ldots, r, j = 1, 2, \ldots s)$ and $(A_i B_j)$ represents the number of persons possessing attributes A_i and B_j, we have $\Sigma A_i = \Sigma B_j = N$ = total frequency.

Here $P(A_i)$ = Probability that a person possesses the attribute

$$P(A_i) = \frac{(A_i)}{N},\ i = 1, 2, \ldots, r \qquad \ldots (1)$$

$P(B_j)$ = Probability that a person possesses the attribute B_j

$$= \frac{(B_j)}{N},\ j = 1, 2, \ldots, s \qquad \ldots (2)$$

$P(A_i B_j)$ = Probability that a person possesses both the attributes A_i and B_j

$$= \frac{(A_i\ B_j)}{N}$$

If $(A_i B_j)_0$ = Expected number of persons possessing both the attributes A_i and B_j

$$= N \times P(A_i B_j)$$

$$= P(A_i) \times P(B_j) \times N$$

$$= \frac{(A_i)\,(B_j)}{N} \text{ using (1) and (2)}$$

Hence $$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \left[\frac{[(A_i B_j) - (A_i B_j)_0]^2}{(A_i B_j)_0} \right]$$

which is distributed as a χ^2 variable with $(v-1) \times (s-1)$ degrees of freedom.

Note (1). For a 2×2 contingency table $\begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}$

$$\chi^2 = \frac{(a+b+c+d)\,(ad-bc)^2}{(a+c)\,(b+d)\,(a+b)\,(c+d)}$$ can be used to save time.

Observe & frequencies	Attributes B_1	Attribute B_2
Attribute A_1	a	b
Attribute A_2	c	d

Expected frequency $= (A_i B_j)_0 = \dfrac{(A_i)\,(B_j)}{N}$

$\therefore$ Expected frequencies are,

	Attribute B_1	Attribute B_2
Attribute A_1	$\dfrac{(a+b)\,(a+c)}{N}$	$\dfrac{(a+b)\,(b+d)}{N}$
Attribute A_2	$\dfrac{(a+c)\,(c+d)}{N}$	$\dfrac{(b+d)\,(c+d)}{N}$

That is, the expected frequency in each cell is

$$= \frac{\text{Product of column total and row total}}{\text{whole total}}$$

Note (2) : If the calculated χ^2 value $<$ table value of χ^2, we accept H_o, namely that the two attributes are independent.

If calculated χ^2 $>$ table value of χ^2, the Hypothesis H_0 is rejected and conclude that the attributes are not independent.

Note (3) : If $\dfrac{a \mid b}{c \mid d}$ is the 2×2 contingency table with two attributes,

$Q = \dfrac{ad - bc}{ad + bc}$ is called the coefficient of association.

If the attributes are independent then $\dfrac{a}{b} = \dfrac{c}{d}$.

Note (4) : YATE's Correction : In a 2×2 table if the frequency of a cell is small say 3 or 4, we make Yate's correction to make χ^2 continuous.

Correction : Calculate *ad* and *bc*, and add 0.5 to both factors of the small product and subtract 0.5 from both factors of the larger product.

Note (5) After Yate's Correction,

$$\chi^2 = \frac{N\left(ad - bc - \frac{1}{2}N\right)^2}{(a+b)(a+c)(c+d)(b+d)}$$

N = total frequency

Example 1.3.c.(1)

For the 2×2 contingency table $\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array}$ the χ^2-test of independence is $\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$ where $N = a + b + c + d$.

Proof : Let the two attributes be A and B. Then the 2×2 contingency table is given below :

Attributes	B	β	Total
A	a	b	$a + b = (A)$
α	c	d	$c + d = (\alpha)$
Total	$(B) = a + c$	$(\beta) = b + d$	$N = a + b + c + d$

Set the null hypothesis H_0 : A and B are independent.

The expected frequencies are given by

$$e(AB) = \frac{(a+b)(a+c)}{N}; \qquad e(A\beta) = \frac{(a+b)(b+d)}{N};$$

$$e(\alpha B) = \frac{(c+d)(a+c)}{N}; \qquad e(\alpha\beta) = \frac{(c+d)(b+d)}{N}$$

$$\chi^2 = \frac{\Sigma(o_i - e_i)^2}{e_i} = \frac{[o(AB) - e(AB)]^2}{e(AB)} + \frac{[o(A\beta) - e(A\beta)]^2}{e(A\beta)}$$

$$+ \frac{[o(\alpha B) - e(\alpha B)]^2}{e(\alpha B)} + \frac{[o(\alpha\beta) - e(\alpha\beta)]^2}{e(\alpha\beta)}$$

Now, $o\ (AB) - e\ (AB) = a - \frac{(a+b)(a+c)}{N} = a - \frac{(a+b)(a+c)}{a+b+c+d}$

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

$$= \frac{ad - bc}{N}$$

Similarly we can get $o(A\ \beta) - e\ (A\beta) = \frac{ad - bc}{N}$;

$o\ (\alpha B) - e\ (\alpha B) = \frac{ad - bc}{N}$; $o\ (\alpha\beta) - e\ (\alpha\beta) = \frac{ad - bc}{N}$

$$\therefore \chi^2 = \frac{(ad - bc)^2}{N^2}\left[\frac{1}{e\ (AB)} + \frac{1}{e\ (A\beta)} + \frac{1}{e(\alpha B)} + \frac{1}{e(\alpha\beta)}\right]$$

$$= \frac{(ad - bc)^2}{N^2}\left[\frac{N}{(a+b)(a+c)} + \frac{N}{(a+b)(b+d)} + \frac{N}{(c+d)(a+c)} + \frac{N}{(c+d)(b+d)}\right]$$

$$= \frac{(ad - bc)^2}{N}\left[\frac{b+d+a+c}{(a+b)(a+c)(b+d)} + \frac{b+d+a+c}{(c+d)(a+c)(b+d)}\right]$$

$$= (ad - bc)^2\left[\frac{c+d+a+b}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$= \left[\frac{N(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}\right]$$

Example 1.3.c.(2)

What are the expected frequencies of 2×2 contingency table $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ **[A.U A/M 2015 R-17]**

Solution : The complete table is

Attributes	B	β	Total
A	a	b	$a+b$
α	c	d	$c+d$
Total	(B) $= a+c$	$(\beta) = b+d$	$a+b+c+d$

The expected frequencies are

$$e(a)=e(A, B) = \frac{(a+b)(a+c)}{a+b+c+d}; \quad e(b)=e(A, \beta) = \frac{(a+b)(b+d)}{a+b+c+d},$$

$$e(c)=e(\alpha, A) = \frac{(c+d)(a+c)}{a+b+c+d}; \quad e(d)=e(\alpha, \beta) = \frac{(c+d)(b+d)}{a+b+c+d}.$$

Example 1.3.c.(3)

Find if there is any association between extravagance in fathers and extravagance in sons from the following data.

	Extravagant father	Miserly father
Extravagant son	327	741
Miserly son	545	234

Determine the coefficient of association also. **[A.U. M/J 2013]**

Solution :

The parameter of interest is χ^2

1. H_0 : Namely that the extreavagance in sons and fathers are not significant.
2. H_1 : Significant.
3. $\alpha = 0.05$, d.f $= (r-1)(s-1) = (2-1)(2-1) = 1$
4. **Table value of χ^2** : 3.841
5. **The test statistic is** $\chi^2 = \frac{(ad-bc)^2 (a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$

i.e., $\chi^2 = \frac{[(327)(234) - (545)(741)]^2 \times (327 + 545 + 741 + 234)}{(872)(975)(1068)(779)} = 230.24$

6. **Conclusion :**

If cal. χ^2 <table χ^2, then we accept H_0 ; otherwise, we reject H_0.

Here, $\chi^2 = 230.24 > 3.841$ solve reject H_0 at 5% level of significance.

$\therefore$ There is dependence between the attributes

7. Coefficient of attributes $= \frac{ad - bc}{ad + bc}$

$= \frac{-327330}{480363} = -0.6814$

Example 1.3.c.(4)

On the basis of information noted below, find out whether the new treatment is comparatively superior to the conventional one.

	Favourable	Non-favourable	Total
Conventional	40	70	90
New	60	30	110
Total	100	100	N = 200

Solution :

The parameter of interest is χ^2

1. H_0 : No difference between the two treatment.
2. H_1 : difference between the two treatment.
3. $\alpha = 0.05$, d.f $= (r - 1)(s - 1) = (2 - 1)(2 - 1) = 1$
4. **Table value of χ^2** : 3.841

5. **The test statistic is** $\chi^2 = \dfrac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$

$$= \frac{(40 \times 30 - 60 \times 70)^2 (200)}{100 \times 100 \times 90 \times 110} = 18.181818$$

8. **Conclusion :**

If cal. χ^2 <table χ^2, then we accept H_0 ; otherwise, we reject H_0.

Here, $\chi^2 = 18.181818 > 3.841$. So, we reject H_0 at 5% level of significance.

We conclude that the treatments are not independent.

Example 1.3.c.(5)

1000 students at college level were graded according to their I.Q. and their economic conditions. What conclusion can you draw from the following data :

Economic conditions	I.Q. Level	
	High	Low
Rich	460	140
Poor	240	160

[A.U. M/J 2013]

Solution :

The parameter of interest is, χ^2

1. H_0 : The given attributes are independent.
2. H_1 : The given attributes are not independent.
3. $\alpha = 0.05$, d.f $= (r - 1)(s - 1) = (2 - 1)(2 - 1) = 1$
4. **Table value of χ^2 :** 3.841
5. **The test statistic is** $\chi^2 = \Sigma \dfrac{(O - E)^2}{E}$

The expected frequencies are calculated using the following formula:

$$\text{Expected Frequency} = \frac{\text{Corresponding Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

$$\text{Expected frequency for } 460 = \frac{600 \times 700}{1000} = 420$$

$$\text{Expected frequency for } 140 = \frac{600 \times 300}{1000} = 180$$

$$\text{Expected frequency for } 240 = \frac{700 \times 400}{1000} = 280$$

$$\text{Expected frequency for } 160 = \frac{300 \times 400}{1000} = 120$$

The Chi-square calculations are given as follows :

O	E	O – E	$(O - E)^2$	$(O - E)^2/E$
460	420	40	1600	3.81
140	180	–40	1600	8.88
240	280	–40	1600	5.714
160	120	40	1600	13.33
				31.7373

$$\therefore \chi^2 = 31.7373$$

6. **Conclusion :**

If cal. χ^2 <table χ^2, then we accept H_0 ; otherwise, we reject H_0.

Here, $\chi^2 = 31.7373 > 3.841$. So, we reject H_0 at 5% level of significance.

We conclude that the attributes I.Q. as Economic conditions are not independent.

Example 1.3.c.(9)

Out of 8000 graduates in a town 800 are females, out of 1600 graduate employees 120 are females. Use χ^2 to determine if any distinction is made in appointment on the basis of sex. Value of χ^2 at 5% level for one degree of freedom is 3.84. **[A.U A/M 2010]**

Solution : Given :

	Male	Female	Total
Graduates in a town	7200	800	8000
Graduate employees	1480	120	1600
Total	8680	920	9600

The parameter of interest is, χ^2

1. H_0 : There is no significant difference between male and female.
2. H_1 : There is significant difference between male and female.
3. $\alpha = 0.05$, d.f $= (r-1)(s-1) = (2-1)(2-1) = 1$
4. **Table value of χ^2** : 3.84
5. **The test statistic is** $\chi^2 = \Sigma \dfrac{(O-E)^2}{E}$

The expected frequencies are calculated using the following formulae.

$$\text{Expected frequency} = \frac{\text{Corresponding Row total} \times \text{Column total}}{\text{Grand total}}$$

$$\text{Expected frequency for } 7200 = \frac{(8000)(8680)}{9600} = 7233.33$$

$$800 = \frac{(8000)(920)}{9600} = 766.67$$

$$1480 = \frac{(1600)\,(8680)}{9600} = 1446.67$$

$$120 = \frac{(1600)\,(920)}{9600} = 153.33$$

O	E	O − E	$\chi^2 = \frac{(O-E)^2}{E}$
7200	7233.33	−33.33	0.1536
800	766.67	33.33	1.4490
1480	1446.67	33.33	0.7679
120	153.33	−33.33	7.2451
			9.6156

6. Conclusion :

If cac. χ^2 <table χ^2, then we accept H_0; otherwise, we reject H_0.

Here, $\chi^2 = 9.6156 > 3.841$.

So, we reject H_0 at 5% level of significance.

Example 1.3.c.(10)

An automobile company gives the following information about age groups and the liking for particular model of car which it plans to introduce. On the basis of this data can it be concluded that the model appeal is independent of the age group. (χ^2 0.05 (3) = 7.815)

Persons who :	Below 20	20 - 39	40 - 59	60 and above
Liked the car :	140	80	40	20
Disliked the car :	60	50	30	80

[A.U A/M 2010]

Solution :