Index

Unit – 5 ⇝ Inferential Statistics – II

Hypothesis Testing for Large Sample – II

Method – 1 ⇝ Test for Single Mean

Test for Single Mean

→ Consider a sample X size n with mean $\overline{x}$ and SD s taken from population with mean μ and SD σ.

→ This test is used to find significant difference between sample **mean $\overline{x}$** and **population mean μ**.

→ In this test,

$$\mathbf{SE(t)} = \begin{cases} \dfrac{\sigma}{\sqrt{n}} & ; \quad \sigma \text{ is known} \\ \dfrac{s}{\sqrt{n}} & ; \quad \sigma \text{ is unknown} \end{cases}$$

→ **Formula for Test Statistics**

- When population SD **σ is known**

$$\mathbf{Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}$$

where, $Q = 1 - P$

- When population SD **σ is not known**

$$\mathbf{Z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}}$$

where, $q = 1 - p$

→ **Confidence Limits**

Confidence Limits $= \overline{x} \pm Z_\alpha \, SE(t)$

where, Z_α = Critical value at level of significance α

Example of Method-1: Test for Single Mean

C	1	Let X be the length of a life of certain computer is approximately normally distributed with mean 800 days and standard deviation 40 days. If a random sample of 30 computers have an average life of 788 days, test the null hypothesis that $\mu \neq 800$ days at 5 % and 15% level of significance. ($\lvert Z_{0.05} \rvert = 1.96$; $\lvert Z_{0.15} \rvert = 1.45$) **Answer: Null hypothesis accepted at 5% and rejected at 15%.**
C	2	The mean IQ of a sample of 1600 children was 99. Is it likely that this was a random sample from a population with mean IQ 100 and SD 15? ($\lvert Z_{0.05} \rvert = 1.96$) **Answer: Sample was not drawn from a population with mean 100 and SD 15.**
C	3	An insurance agent has claimed that the average age of policy-holders who insure through him is less than the average for all agent, which is 30.5 years. This sample was drawn from a sample whose mean is 28.8 km and a standard deviation of 6.35 km. Test the significance at 0.05 level. ($Z_{0.05} = -1.645$) **Answer: An insurance agent's claim is valid.**
C	4	A college claims that its average class size is 35 students. A random sample of 64 students from class has a mean of 37 with a standard deviation of 6. Test at the $\alpha = 0.05$ level of significance if the claimed value is too low. ($Z_{0.05} = 1.645$) **Answer: The true mean class size is likely to be more than 35.**

Method – 2 ⇝ Test for Difference of Means

Test for Difference of Means

→ Consider two samples X_1 and X_2 of sizes n_1 & n_2, mean $\overline{x_1}$ & $\overline{x_2}$ and SD s_1 and s_2 respectively taken from two different population of sizes N_1 & N_2, mean $\overline{\mu_1}$ & $\overline{\mu_2}$ and SD σ_1 and σ_2.

→ This test is used to find the significant difference between **sample means $\overline{x_1}$ & $\overline{x_2}$**.

→ In this test,

$$\mathbf{SE(t)} = \begin{cases} \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} & ; \quad \sigma \text{ is known} \\[2ex] \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} & ; \quad \sigma \text{ is unknown} \end{cases}$$

→ **Formula for Test Statistics**

- When population SD **σ is known**

$$\mathbf{Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

where, $Q = 1 - P$

- When population SD **σ is not known**

$$\mathbf{Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

where, $q = 1 - p$

→ **Confidence Limits**

Confidence Limits $= (\overline{x_1} - \overline{x_2}) \pm Z_\alpha\, SE(t)$

where, Z_α = Critical value at level of significance α

Example of Method-2: Test for Difference of Means

C	1	Test the significance of the difference between the means of two normal population with the same standard deviation from the following data: <table><tr><td></td><td>Size</td><td>Mean</td><td>SD</td></tr><tr><td>Sample I</td><td>100</td><td>64</td><td>6</td></tr><tr><td>Sample II</td><td>200</td><td>67</td><td>8</td></tr></table>($\lvert Z_{0.05} \rvert = 1.96$) **Answer: There is significant difference between the means.**
C	2	In a random sample of 100 light bulbs manufactured by a company A, the mean lifetime of light bulb is 1190 hours with standard deviation of 90 hours. Also, in a random sample of 75 light bulbs manufactured by company B, the mean lifetime of light bulb is 1230 hours with standard deviation of 120 hours. Is there a difference between the mean lifetime of the two brands of light bulbs at a significance level of 0.05 and 0.01? ($\lvert Z_{0.05} \rvert = 1.96$; $\lvert z_{0.01} \rvert = 2.58$) **Answer: There is difference between the mean lifetimes at 5%.** **There is no difference between the mean lifetimes at 1%.**
C	3	The means of simple samples of size 1000 and 2000 are 67.5 and 68 cm respectively. Can the sample be regarded as drawn from the same proportion of SD 2.5 cm. ($\lvert Z_{0.05} \rvert = 1.96$) **Answer: Samples are not drawn from population having SD 2.5 cm.**

C	4	A company A manufactured tube lights and claims that its tube lights are superior than its main competitor company B. The study showed that a sample of 40 tube lights manufactured by company A has a mean lifetime of 647 hours of continuous use with a standard deviation of 27 hours, while a samle of 40 tube lights manufactured by company had a mean lifetime 638 hours of continuous use with a standard deviation of 31 hours. Does this substantiate the claim of company A that their tube lights are superior than manufactured by company B at 0.05 and 0.01 level of significance? ($Z_{0.05} = 1.645$; $Z_{0.01} = 2.33$) **Answer: The claim of company A is not valid at 5%.** **The claim of company A is not valid at 1%.**
C	5	The average of marks scored by 32 boys is 72 with standard deviation 8 while that of 36 girls is 70 with standard deviation 6. Test at 1% level of significance whether the boys perform better than the girls. ($Z_{0.05} = 1.645$) **Answer: The boys do not perform better than the girls.**

Method – 3 ⇝ Test for Difference of Standard Deviation

Test for Difference of Standard Deviation

→ Consider two samples X_1 and X_2 of sizes n_1 & n_2, and SD s_1 and s_2 respectively taken from two different population of sizes N_1 & N_2, and SD σ_1 and σ_2.

→ This test is used to find the significant difference between **sample SDs s_1 & s_2**.

→ In this test,

$$\mathbf{SE(t)} = \begin{cases} \sqrt{\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}} & ; \quad \sigma \text{ is known} \\[2ex] \sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}} & ; \quad \sigma \text{ is unknown} \end{cases}$$

→ **Formula for Test Statistics**

- When population SD **σ is known**

$$\mathbf{Z} = \frac{\mathbf{s_1 - s_2}}{\sqrt{\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}}}$$

- When population SD **σ is not known**

$$\mathbf{Z} = \frac{\mathbf{s_1 - s_2}}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}}$$

→ **Confidence Limits**

Confidence Limits $= (s_1 - s_2) \pm Z_\alpha \, SE(t)$

where, Z_α = Critical value at level of significance α

Example of Method-3: Test for Difference of Standard Deviations

<table>
<tr><td>C</td><td>1</td><td>The SD of a random sample of 900 members is 4.6 and that of another independent sample of 1600 members is 4.8. Examine if the two samples could have been drawn from a population with SD 4? ($|Z_{0.05}| = 1.96$)

Answer: Two samples could have been drawn from a population with SD 4.</td></tr>
<tr><td>C</td><td>2</td><td>Random samples drawn from two countries gave the following data relating to the heights of adult males:
<table>
<tr><td></td><td>Country A</td><td>Country B</td></tr>
<tr><td>Standard Deviation</td><td>2.58</td><td>2.5</td></tr>
<tr><td>Number in samples</td><td>1000</td><td>1200</td></tr>
</table>
Is the difference between the standard deviation significant?
($|Z_{0.05}| = 1.96$)

Answer: There is no significant difference between sample SDs.</td></tr>
</table>

Introduction

→ If the sample are large $(n > 30)$ then the sampling distribution of a statistics is normal. But if the samples are small $(n < 30)$ then method that we have seen in previous unit does not hold good.

→ For estimating of the parameter as well as for testing a hypothesis, following distributions are used:

- Student's t – distribution
- Snedecor's F – distribution
- Chi – Square distribution

Degree of Freedom

→ The number of independent pieces of information used to calculate a statistic is known as **degrees of freedom(df)**.

→ In other words, they are the number of values that are able to be changed in a data set.

→ For Example:

- Let two numbers: x, y. The mean of those numbers is m.
- In this data set of three variables $(x, \ y, \ m)$, if you choose the values of any two variables, the third one is already determined.
- Let, $x = 2, \ \ y = 4$. You can't choose any mean you like as it's already determined as below:

$$m = \frac{x + y}{2} = 3$$

- Similarly, If you take, $x = 3, \ \ m = 6$, then y's value is automatically set. It is not free to change.

$$m = \frac{x + y}{2}$$

$$\Rightarrow 6 = \frac{3 + y}{2}$$

$$\Rightarrow y = 9$$

- Any time you assign some two values, the third has no **"freedom to change"**. Hence, there are **2** degrees of freedom in this example.

Working Rule for Hypothesis Testing

→ **Step 1:** Set up null hypothesis H_0.

→ **Step 2:** Set up alternative hypothesis H_1.

→ **Step 3:** Set up level of significance α.

→ **Step 4:** Apply test statistic.

→ **Step 5:** Find appropriate degree of freedom and set up critical region. (Given in data OR to find from statistical tabular table).

→ **Step 6:** Conclusion.

- Compare the computed value of $\mathbf{t}$ with critical value $\mathbf{t}_{\alpha,v}$.
- If $|\ \mathbf{t}\ | > |\ \mathbf{t}_{\alpha,v}\ |$, we **reject H_0** and conclude that there is significant difference.
- If $|\ \mathbf{t}\ | < |\ \mathbf{t}_{\alpha,v}\ |$, we **accept H_0** and conclude that there is no significant difference.

Hypothesis Testing for Small Sample

Method – 4 ⇝ t - Test for Single Mean

t - Test for Single Mean

→ Consider a sample X of size n with mean $\overline{x}$ taken from normal population with mean μ of size N.

→ This test is used to find the significant difference between **mean of sample $\overline{\mathbf{x}}$ & mean of population μ** when variance of the population is unknown.

→ In this test,

$$\mathbf{SE(t)} = \mathbf{f(x)} = \begin{cases} \dfrac{s}{\sqrt{n-1}} & ; \quad \text{when SD is given in data} \\[2ex] \dfrac{S}{\sqrt{n}} & ; \quad \text{when SD is not given in data} \end{cases}$$

$$\mathbf{df}\ (\ \mathbf{v}\ ) = n - 1$$

$$\text{Where, } s = \text{sample SD} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$S = \text{unbiased esitimation of sample SD} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

→ Relationship between sample SD " s " & estimates of SD " S " is $\mathbf{n \cdot s^2 = (\ n - 1\ ) \cdot S^2}$

$$\Rightarrow \frac{s^2}{n-1} = \frac{S^2}{n}$$

$$\Rightarrow \frac{s}{\sqrt{n-1}} = \frac{S}{\sqrt{n}}$$

→ **Formula for Test Statistics**

$$\mathbf{t = \frac{\overline{x} - \mu}{SE(t)}}$$

→ **Confidence Limits**

Confidence Limits $\mathbf{= \overline{x} \pm t_{\alpha,v}\ SE(t)}$

where, $t_{\alpha,v}$ = Critical value at level of significance α at degree of freedom (v).

Example of Method-4: t - Test for Single Mean

C	1	A random sample of size 16 has 53 as mean. The sum of squares of the derivation from mean is 135. Can this sample be regarded as taken from population having 56 as mean? $\left(\lvert t_{0.01,15} \rvert = 2.1314 \right)$ **Answer: The sample mean has not come from a population mean 56.**
C	2	The heights of 10 males of a given locality are found to be 175, 168, 155, 170, 152, 170, 175, 160, 160 and 165 cm. Based on this sample, find the 95% confidence limits for the heights of males in that locality. $\left(\lvert t_{0.05,9} \rvert = 2.2622 \right)$ **Answer: $\mathbf{159.2671}$ and $\mathbf{170.7329}$**
C	3	Ten individuals were chosen random from a normal population and their heights were found to be in inches $63, 63, 66, 67, 68, 69, 70, 70, 71$ and 71. Test the hypothesis that the mean height of the population is 66 inches. $\left(\lvert t_{0.05,9} \rvert = 2.2622 \right)$ **Answer: There is no significant difference between population mean and sample mean.**
C	4	A soap manufacturing company was distributing a particular brand of soap through large number of retail shops. Before a heavy advertisement campaign, the mean sales per week per shop was 140 dozen. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozen with standard deviation 16. Can you consider the advertisement effective? $\left(t_{0.05,25} = 1.7081 \right)$ **Answer: The advertisement is effective.**

Method – 5 ⇝ t - Test for Difference of Means

t - Test for Difference of Means

→ Consider two samples X_1 and X_2 of sizes n_1 & n_2 and of mean $\overline{x}$ & $\overline{y}$ respectively taken from two different population of sizes N_1 & N_2 and mean μ_X & μ_Y.

→ This test is used to find the significant difference between **sample means $\overline{x}$ & $\overline{y}$** under the assumption that the population variance are equal. $(\sigma_X = \sigma_Y = \sigma)$.

→ In this test,

$$\mathbf{SE(t)} = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\mathbf{df}\ (\mathbf{v}) = n_1 + n_2 - 2$$

Where,

$$S = \sqrt{\frac{1}{n_1 + n_2 - 2} \cdot \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} (y_i - \overline{y})^2\right]} \quad ; \text{ when SD is not given in data}$$

$$= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \quad ; \text{ when SD is given in data}$$

→ **Formula for Test Statistics**

$$\mathbf{t} = \frac{\overline{\mathbf{x}} - \overline{\mathbf{y}}}{\mathbf{SE(t)}}$$

→ **Confidence Limits**

Confidence Limits $= (\overline{\mathbf{x}} - \overline{\mathbf{y}}) \pm \mathbf{t}_{\alpha,v}\, \mathbf{SE(t)}$

where, $t_{\alpha,v}$ = Critical value at level of significance α at degree of freedom (v).

Example of Method-5: t - Test for Difference of Means

C 1

Two types of batteries are tested for their length of life and the following data are obtained:

	No. of samples	Mean life in hours	Variance
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means? Find 95% confidence limits for the difference in means. ($|t_{0.05,15}| = 2.1314$)

Answer: There is significant difference in the two means.

Confidence Limits: -27.3411 and -52.65889

C 2

The following data represents the biological values of protein from cow's milk and buffalo's milk at a certain level:

Cow's milk	1.82	2.02	1.88	1.61	1.81	1.54
Buffalo's milk	2.00	1.83	1.86	2.03	2.19	1.88

Examine if the average values of protein in the two samples significantly differ. ($|t_{0.05,10}| = 2.2281$)

Answer: There is no significant difference in average values of proteins in two milk samples.

C 3

The mean height and SD height of 8 randomly chosen soldiers are 166.9 cm and 2.29 cm respectively. The corresponding values of 6 randomly chosen sailors are 107.3 cm and 8.50 cm respectively. Based on this data, can we conclude that soldiers are, in general, shorter than sailors?
($t_{0.05,12} = -1.7823$)

Answer: The soldiers are not shorter than sailors.

C	4	Samples of two types of electric bulbs were tested for length of life and the following data were obtained. | | Size | Mean | SD | | Sample 1 | 8 | 1234 | 36 hr. | | Sample 2 | 7 | 1036 | 40 hr. | Is the difference in the means sufficient to warrant that type 1 bulbs are superior to type 2 bulbs? ($t_{0.05,13} = 1.7709$) **Answer: The type 1 bulbs are superior to type 2 bulbs.**

Method – 6 ⇝ t - Test for Correlation Coefficient

t - Test for Correlation Coefficient

→ Consider two samples X_1 and X_2 of sizes n_1 & n_2 with correlation coefficient r taken from population with correlation coefficient ρ.

→ This test is used to find whether the sample correlation is significant of any correlation in the population or not.

→ In this test,

$$\mathbf{SE(t)} = \frac{1 - r^2}{\sqrt{n}}$$

$$\mathbf{df}\ (\ \mathbf{v}\ ) = n - 2$$

→ **Formula for Test Statistics**

$$\mathbf{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

→ **Confidence Limits**

$$\mathbf{Confidence\ Limits = r \pm t_{\alpha,v}\ SE(t)}$$

where, $t_{\alpha,v}$ = Critical value at level of significance α at degree of freedom (v).

Example of Method-6: t - Test for Correlation Coefficient

C	1	A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0.6. Is this significant of correlation in the population? ($\lvert t_{0.05,25} \rvert = 2.0595$) **Answer: The sample population is correlated.**
C	2	The correlation coefficient between income and food expenditure for sample of 7 household from a low-income group is 0.9. Using 1% level of significance, test whether the correlation coefficient between incomes and food expenditure is positive. Assume that the population of both variables are normally distributed. ($t_{0.01,5} = 3.3649$) **Answer: There is correlation between incomes and food expenditure.**

Method – 7 ⇝ F – Test for Ratio of Variances

Test for F – Test for Ration of Variances

→ Consider two samples X_1 and X_2 of sizes n_1 & n_2, mean $\overline{x_1}$ & $\overline{x_2}$ respectively taken from two different population of sizes N_1 & N_2, mean μ and SD σ.

→ This test is used to find the significance of the difference between **population standard deviations** σ_1 and σ_2 using unbiased estimates of SDs **S_1 and S_2**.

→ In this test,

df (**v**) $= n - 1$

→ **Formula for Test Statistics**

- If $\mathbf{S_1 > S_2}$,

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

- If $\mathbf{S_2 > S_1}$,

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

$$\text{where, } S_1 = \sqrt{\frac{1}{n_1 - 1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$S_2 = \sqrt{\frac{1}{n_2 - 1} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}$$

→ Relationship between sample SD " s " & estimates of SD " S " is $\mathbf{n \cdot s^2 = (n - 1) \cdot S^2}$

$$\Rightarrow S^2 = \frac{n \cdot s^2}{n - 1}$$

$$\text{So, } {S_1}^2 = \frac{n_1 \cdot {s_1}^2}{n_1 - 1} \quad ; \quad {S_2}^2 = \frac{n_2 \cdot {s_2}^2}{n_2 - 1}$$

Example of Method-7: F - Test for Ratio of Variances

C 1

In a laboratory experiment two samples gave the following results:

Sample no.	Size	Mean	Variance
I	10	15	90
II	12	14	108

Test the equality of sample variances at 5 % level of significance.

$(F_{0.05}(11, 9) = 3.10)$

Answer: The two population have the same variances.

C 2

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

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

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

$(F_{0.05}(6, 4) = 6.16)$

Answer: The variances of time distribution in a population from which samples are drawn do not differ significantly.

C 3

Two random samples gave the following data:

Sample no.	Size	Mean	Variance
I	16	9.6	40
II	25	16.5	42

Can we conclude that the two samples have been drawn from the same normal population? $\left(F_{0.05}(24, 15) = 2.29 \ ; \left| t_{0.05,39} \right| = 2.0227 \right)$

Answer: The two samples are not drawn from the same normal population.

C	4	Two nicotine contents in two random samples of tobacco are given below:

Sample I	21	24	25	26	27	-
Sample II	22	27	28	30	31	36

Can we say that two samples came from the same population?

$\left(F_{0.05}(5,4) = 6.26 \ ; \ \left| t_{0.05,9} \right| = 2.2622 \right)$

Answer: Two samples came from the same population.

Chi – Square Test: Introduction

→ The chi–square (χ^2) test is a useful measure of comparing experimentally obtained results with those expected theoretically and based on hypothesis.

→ Symbol χ^2 is read as "**Ky Square**".

→ It is used as a test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared.

→ The magnitude of discrepancy between observed and theoretical frequencies is given by the quantity χ^2.

→ If $\chi^2 = 0$, the observed and expected frequencies completely coincides. As the value of χ^2 increases, the discrepancy between the observed and theoretical frequency decreases.

→ If $f_{o_1}, f_{o_2}, \dots, f_{o_n}$ be a set of observed frequencies and $f_{e_1}, f_{e_2}, \dots, f_{e_n}$ be the corresponding set of expected frequencies χ^2 then it is defined by,

$$\chi^2 = \frac{(f_{o_1} - f_{e_1})^2}{f_{e_1}} + \frac{(f_{o_2} - f_{e_2})^2}{f_{e_2}} + \cdots + \frac{(f_{o_n} - f_{e_n})^2}{f_{e_n}} = \sum_{i=1}^{n} \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}}$$

→ **Conditions for validity of χ^2 test:**

- The sample observations should be independent.
- The total frequency $N = \sum f_i$ should be reasonably large, say, greater than 50.
- Each expected frequency $f_{e_i} \geq 5$. If not, then it is pooled with preceding or succeeding frequency so that the pooled frequency is more than 5. In this case, we need to adjust degree of freedom lost in pooling.

→ **Applications of χ^2 test:**

- Goodness of Fit
- Independence of Attributes

Working Rule for Chi- Square Test

→ **Step 1:** Set up null hypothesis H_0.

→ **Step 2:** Set up alternative hypothesis H_1.

→ **Step 3:** Set up level of significance α.

→ **Step 4:** Find appropriate f_e and apply test statistic.

→ **Step 5:** Find appropriate degree of freedom and set up critical region. (Given in data OR to find from statistical tabular table).

→ **Step 6:** Conclusion.

- Compare the computed value of χ^2 with critical value $\chi_{\alpha,v}$.
- If $\chi^2 > \chi_{\alpha,v}$, we **reject H_0** and conclude that there is significant difference.
- If $\chi^2 < \chi_{\alpha,v}$, we **accept H_0** and conclude that there is no significant difference.

Method – 8 ⇝ Chi – Square Test: Goodness of Fit

Chi-Square Test: Goodness of Fit

→ The values of χ^2 is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not.

→ It is also used test to check fitting a set of observations to given distribution or not.

→ **Degree of freedom (df)**

- If the data is given in a series of n numbers, $\mathbf{v = n - 1}$
- In case of binomial distribution, $\mathbf{v = n - 1}$
- In case of poisson distribution, $\mathbf{v = n - 2}$
- In case of normal distribution, $\mathbf{v = n - 3}$

→ **Expected Frequency ($\mathbf{f_e}$)**

- If the data is given in a series of n numbers,

$$\mathbf{f_e = mean = \frac{\sum x_i}{n}}$$

Where, $n = \text{Total observation}$

- In case of binomial distribution,

$$\mathbf{f_e = N \cdot p(x)}$$

Where, $p(x) = {}_nC_x \cdot p^x \cdot q^{n-x}$

$x = \text{Observed data } (x_i)$

$n = \text{Total observation}$

$N = \text{Sum of observed frequency}$

- In case of poisson distribution,

$$\mathbf{f_e = N \cdot p(x)}$$

Where, $p(x) = \dfrac{e^{-\lambda} \cdot \lambda^x}{x!}$

$x = \text{Observed data } (x_i)$

$N = \text{Sum of observed frequency}$

$\lambda = \dfrac{\sum x_i}{N}$

- In case of normal distribution,

$$\mathbf{f_e = N \cdot \text{Difference of area under the curve up to } Z_{x_i} \text{ and } Z_{x_{i+1}}}$$

Where, $Z_x = \dfrac{\overline{x} - \mu}{\sigma}$

$\overline{x}$ = Sample Mean (Given in Data OR Manually calculate from data)

μ = Population Mean (Given in Data)

σ = Population SD (Given in Data)

→ **Formula for Test of Significance**

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

Example of Method-8: Chi – Square Test: Goodness of Fit

<table>
<tr><td>C</td><td>1</td><td>Suppose that a die is tossed 120 times and the recorded data is as follows:

Face Observed(x)	1	2	3	4	5	6
Frequency	20	22	17	18	19	24

Test the hypothesis that the die is unbiased at $\alpha = 0.05$.
$\left(\chi^2{}_{0.05,5} = 11.070 \right)$

Answer: The die is unbiased.</td></tr>
<tr><td>C</td><td>2</td><td>Theory predicts that the proportion of beans in the four group A, B, C, D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Do the experimental results support the theory? $\left(\chi^2{}_{0.05,3} = 7.815 \right)$

Answer: Experimental results support the theory.</td></tr>
</table>

C	3	Records taken of the number of male and female births in 830 families having four children are as follows: <table><tr><td>No. of male births</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td></tr><tr><td>No. of female births</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td></tr><tr><td>No. of families</td><td>32</td><td>178</td><td>290</td><td>236</td><td>94</td></tr></table>Test whether data are consistent with hypothesis that the binomial law holds and the chance of male birth is equal to that of female birth, namely. $\left(\chi^2{}_{0.05,4} = 9.488 \right)$ **Answer: The data are not consistence with the hypothesis.**
C	4	Suppose that during 400 five-minute intervals the air-traffic control of an airport received $0, 1, 2, \ldots$ or 13 radio messages with respective frequencies of $3,\ 15,\ 47,\ 76,\ 68,\ 74,\ 46,\ 39,\ 15,\ 9,\ 5,\ 2,\ 0$ and 1. Test at 0.05 level of significance, the hypothesis that the number of radio messages received during 5 minute interval follows Poisson distribution with $\lambda = 4.6$. $\left(\chi^2{}_{0.05,8} = 15.507 \right)$ **Answer: Poisson distribution with $\boldsymbol{\lambda = 4.6}$ provides a good fit.**
C	5	The following table indicates (a) the frequencies of a given distribution with (b) the frequencies of the normal distribution having the same mean, standard deviation and the total frequency as in (a). <table><tr><td>(a)</td><td>1</td><td>5</td><td>20</td><td>28</td><td>42</td><td>22</td><td>15</td><td>5</td><td>2</td></tr><tr><td>(b)</td><td>1</td><td>6</td><td>18</td><td>25</td><td>40</td><td>25</td><td>18</td><td>6</td><td>1</td></tr></table>Apply the χ^2 - test of goodness of fit. $\left(\chi^2{}_{0.05,4} = 9.488 \right)$ **Answer: This normal distribution provides a good fit.**
C	6	Fit the equation of the best fitting normal curve to the following data: <table><tr><td>X</td><td>135</td><td>145</td><td>155</td><td>165</td><td>175</td><td>185</td><td>195</td><td>205</td></tr><tr><td>f</td><td>1</td><td>6</td><td>18</td><td>25</td><td>40</td><td>25</td><td>18</td><td>6</td></tr></table>Compare the theoretical and observed frequencies. Using χ^2 - test find goodness of fit. Given that $\mu = 165.6$ and $\sigma = 15.02$. $\left(\chi^2{}_{0.05,2} = 5.991 \right)$ **Answer: This normal distribution provides a good fit.**

Method – 9 ⇝ Chi – Square Test: Independence of Attributes

Chi – Square Test: Independence of Attributes

→ In statistics, sometimes we have to deal with attributes or qualitative characters, which cannot be measured accurately, although items can be divided into two or more categories with respect to the attributes.

→ Let A and B be two attributes of the population. A can be divided into m categories $A_1, A_2, \dots, A_m$ and B can be divided into n categories $B_1, B_2, \dots B_n$.

→ The data can be shown in the form of a two-way table with m rows and n columns, as in a bivariate frequency distribution. This two-way frequency table for attributes is known as $m \times n$ **contingency table**.

→ The frequency of observations belonging to both categories A_i and B_j simultaneously is shown in the cell at i^{th} row and j^{th} column and denoted by $(A_i\, B_j)$. Similarly $(A_i)$and (B_j) denote the frequency of items belonging to categories A_i and B_j respectively and N, the total frequency.

Contingency Table					
Attributes	B_1	B_2	B_3	B_4	Total
A_1	$(A_1\, B_1)$	$(A_1\, B_2)$	$(A_1\, B_3)$	$(A_1\, B_4)$	(A_1)
A_2	$(A_2\, B_1)$	$(A_2\, B_2)$	$(A_2\, B_3)$	$(A_2\, B_4)$	(A_2)
A_3	$(A_3\, B_1)$	$(A_3\, B_2)$	$(A_3\, B_3)$	$(A_3\, B_4)$	(A_3)
Total	(B_1)	(B_2)	(B_3)	(B_4)	N

→ **Degree of freedom (df)**:

$$\mathbf{df = v = (\text{No. of row} - 1) \times (\text{No. of column} - 1)}$$

→ **Expected Frequency (f_e) for ($A_i\, B_j$)**

$$\mathbf{f_{e\,(A_i\, B_j)} = \frac{(\text{Total of } i^{th} \text{ column})\,(\text{Total of } j^{th} \text{ row})}{\text{Total of all frequency}} = \frac{(A_i)\,(B_j)}{N}}$$

→ **Formula for Test of Significance**

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

Example of Method-9: Chi – Square Test: Independence of Attributes

<table>
<tr><td>C</td><td>1</td><td>1000 students at college level were graded according to their IQ and the economic conditions of their home. Use χ^2 test to find out whether there is any association between condition at home and IQ. $\left(\chi^2_{0.05,1} = 3.841\right)$
<table>
<tr><td rowspan="2">Economic Condition</td><td colspan="2">IQ</td></tr>
<tr><td>High</td><td>Low</td></tr>
<tr><td>Rich</td><td>460</td><td>140</td></tr>
<tr><td>Poor</td><td>240</td><td>160</td></tr>
</table>
Answer: Two attributes are not associated.</td></tr>
<tr><td>C</td><td>2</td><td>Test the hypothesis at 5% level of significance that the presence or absence of hypertension is independent of smoking habits from the following data of 180 persons. $\left(\chi^2_{0.05,2} = 5.991\right)$
<table>
<tr><td></td><td>Non smokers</td><td>Moderate smokers</td><td>Heavy smokers</td></tr>
<tr><td>HT</td><td>21</td><td>36</td><td>30</td></tr>
<tr><td>No HT</td><td>48</td><td>26</td><td>19</td></tr>
</table>
Answer: Hypertension and smoking habits are not independent.</td></tr>
</table>

* * * * * **End of the Unit** * * * * *