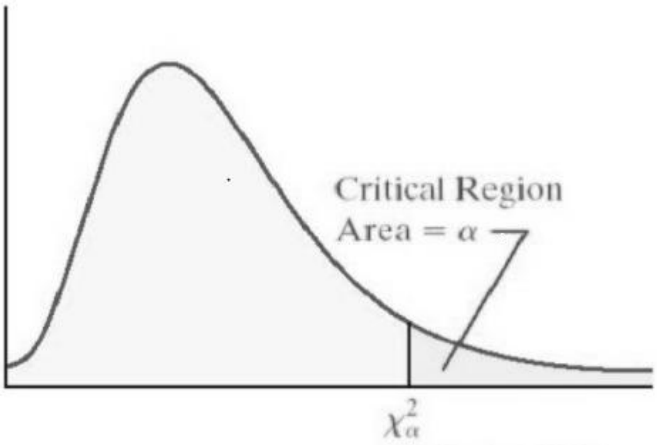
Tests of Significance Based on Chi-Square (χ^2)

7.1 Chi-Square (χ^2) Test

Tests like t , $F$and Z are based on the assumption that the samples are drawn from a normally distributed population. As these tests require assumptions about the type of population or parameters, these tests are called 'parametric tests'.

Sometimes it is unrealistic to make any rigid assumption about the distribution of the population from which samples are drawn. Studies came out with Chi-Square (read as Ki-square) tests which provide non-parametric approach for testing of goodness of fit and independence of attributes. Although it assumes population to be normal while assessing population variance with sample parameters; thus it is parametric in this case.



Chi-square (χ^2) is a right tailed test and describes the magnitude of discrepancy between theory and observation. If $\chi^2 = 0$; the observed and expected frequencies completely coincide, $\therefore$ χ^2 provides a measure of correspondence between theory and observations. It is one of the simplest and most general tests and can be used to perform

1. test of significance of sample variance
2. test of goodness of fit
3. test of independence of attributes in a contingency table

7.2 Chi-Square (χ^2) Test for Testing Significance of Sample Variance

Chi-Square test can be used to check the population variance for a specified value and also if it needs a revision (if population variance is already given), looking into sample parameters.

If $x_1, x_2, \cdots, x_n$ be a random sample of size $n \leq 30$; with variance s^2, from a normal population with given variance σ^2, then under null hypothesis that population variance is unchanged, the statistic Chi- Square is defined by: $\chi^2 = \sum \frac{(x-\overline{x})^2}{\sigma^2} = \frac{1}{\sigma^2}\sum(x-\overline{x})^2$ or $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

where $\overline{x} = \frac{\sum x}{n}$, $s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$ $\Rightarrow \sum(x-\overline{x})^2 = (n-1)s^2$

$\therefore \chi^2$ is unbiased estimator of testing population variance with degrees of $\text{freedom}(\nu) = n-1$.

Example1 The standard deviation of a certain dimension of articles produced by a machine is 7.5 over a long period. A random sample of 25 articles gave a standard deviation of 10.0. Is it justified to conclude that variability has increased?

Value of χ^2 for 24 degrees of freedom at 5% significance level is 36.415

Solution: Let H_0: Population variance $(\sigma^2) = 7.5$ has not changed.

H_1: Variance has increased

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{24\ (10)^2}{(7.5)^2} = 42.67$$

Table value of χ^2 for 24 degrees of freedom at 5% level is 36.415

Calculated value of χ^2 is greater than table value. $\therefore$ H_0 is rejected and it is justified to conclude that variability has increased.

Example2 Eleven measurements of the same object on an instrument; at different times are given by: 2.5, 2.3, 2.4, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5 and 2.3.

Test at 1% level of significance that variance of the instrument is not more than 0.16.

Solution: Let H_0: Population variance $(\sigma^2) = 0.16$

H_1: $\sigma^2 > 0.16$

$\chi^2 = \frac{1}{\sigma^2}\sum(x-\overline{x})^2$, where $\overline{x} = \frac{\sum x}{n}$

Here $\overline{x} = \frac{2.5+\ 2.3+\ 2.4+2.5+2.7+\ 2.5+\ 2.6+\ 2.6+2.7+2.5+\ 2.3}{11} = 2.51$

$$\therefore \sum(x-\overline{x})^2 = 2(2.3-2.51)^2 + (2.4-2.51)^2 + 4(2.5-2.51)^2 + 2(2.6-2.51)^2 + 2(2.7-2.51)^2 = 0.1891$$

$$\therefore \chi^2 = \frac{1}{\sigma^2}\sum(x-\overline{x})^2 = \frac{0.1891}{0.16} = 1.182$$

Table value of χ^2 for 10 degrees of freedom at 1% level is 23.209

Calculated value of χ^2 is greater than table value. $\therefore$ H_0 is rejected and it is justified to conclude that variability has increased.

7.3 Chi-Square (χ^2) Test for Testing Goodness of Fit

If O_i $(i = 1,2, \ldots, n)$ be a set of observed (experimental) frequencies and E_i $(i = 1,2, \ldots, n)$ be the corresponding set of expected (theoretical) frequencies,

then $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$; with degrees of freedom $(\nu) = n - 1$.

There are some underlying conditions for applying χ^2 test such as:

1. Sum of frequencies should be large (at least 50)
2. No theoretical cell-frequencies should be small, if small (less than 5) theoretical frequencies occur, regrouping of two or more cells must be done before calculating
   $(O_i - E_i)$. Degree of freedom is determined by the number of classes after regrouping.

Example3 Apply χ^2 test of goodness to fit for the following data:

Observed frequency:	1	5	20	28	42	22	15	5	2
Theoretical Frequency:	1	6	18	25	40	25	18	6	1

Value of $\chi^2$for 6, 7, 8 degrees of freedom at 5% significance level are 12.592, 14.067 and 15.507 respectively.

Solution: Since first two and the last two cells frequencies are smaller than prescribed, regrouping the data to 7 cells:

O_i	6	20	28	42	22	15	7
E_i:	7	18	25	40	25	18	7

Degrees of freedom $(\nu) = 7 - 1 = 6$

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

$$= \frac{(6-7)^2}{7} + \frac{(20-18)^2}{18} + \frac{(28-25)^2}{25} + \frac{(42-40)^2}{40} + \frac{(22-25)^2}{25} + \frac{(15-18)^2}{18} + \frac{(7-7)^2}{7} = 1.685$$

Table value of χ^2 for 6 degrees of freedom at 5% level = 12.592

Calculated value of χ^2 is much less than table value $\therefore$ the fit is very good.

Example4 In experimental breeding, Mendal got the following frequencies of seeds:

315 round and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green, total 556. Theory predicts that the frequencies should be proportional in the order 9:3:3:1. Examine the correspondence between theory and experiments.

(Value of $\chi^2$for 3 degrees of freedom at 5% significance level = 7.815)

Solution: Expected (theoretical) frequencies are:

$\frac{9}{16} \times 556 \approx 313, \frac{3}{16} \times 556 \approx 104, \frac{3}{16} \times 556 \approx 104, \frac{1}{16} \times 556 \approx 35$

Comparing observed and expected frequencies:

O_i:	315	101	108	32
E_i :	313	104	104	35

Degrees of freedom $(\nu) = 4 - 1 = 3$

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

$$= \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35} = 0.51$$

Table value of χ^2 for 3 degrees of freedom at 5% level = 7.815

Calculated value of χ^2 is much less than table value, $\therefore$ there is much correspondence between theory and experiment.

Example5 A die is thrown 246 times and results of these throws are given as:

Number on the die	1	2	3	4	5	6
Frequency	32	35	59	57	39	24

Find the value of χ^2on the hypothesis that the die is unbiased.

Solution: Let H_0: die is unbiased.

Expected frequency of each number for an unbiased die is $\frac{246}{6} = 41$

Comparing observed and expected frequencies:

O_i	32	35	59	57	39	24
E_i:	41	41	41	41	41	41

Degrees of freedom $(\nu) = 6 - 1 = 5$

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

$$= \frac{(32-41)^2}{41} + \frac{(35-41)^2}{41} + \frac{(59-41)^2}{41} + \frac{(57-41)^2}{41} + \frac{(39-41)^2}{41} + \frac{(24-41)^2}{41} = 24.15$$

Table value of χ^2 for 5 degrees of freedom at 5% level = 11.07

Calculated value of χ^2 is greater than table value, $\therefore$ the die can be concluded to be biased one.

7.4 Chi-Square (χ^2) Test for Testing Independence of Attributes

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables.

Consider two attributes A and B to be tested for independence with categories A_i , $i = 1,2, \dots, m$ and B_j , $j = 1,2, \dots, n$ respectively.

Attribute	$B \rightarrow$					Total
A $\downarrow$		B_1	B_2		B_n	
	A_1	O_{11}	O_{12}	$\cdots$	O_{1n}	$\sum_{j=1}^{n} O_{1j}$
	A_2	O_{21}	O_{22}		O_{2n}	$\sum_{j=1}^{n} O_{2j}$
	$\vdots$					
	A_m	O_{m1}	O_{m2}		O_{mn}	$\sum_{j=1}^{n} O_{mj}$
Total		$\sum_{i=1}^{m} O_{i1}$	$\sum_{i=1}^{m} O_{i2}$		$\sum_{i=1}^{m} O_{in}$	$N = \sum_{i.j} O_{ij}$

Here O_{ij} is observed frequency for attributes A_i B_j in the given contingency table.

Expected frequency E_{ij} is calculated as shown:

$$E_{11} = \frac{(\sum_{j=1}^{n} O_{1j})(\sum_{i=1}^{m} O_{i1})}{N}, \; E_{12} = \frac{(\sum_{j=1}^{n} O_{1j})(\sum_{i=1}^{m} O_{i2})}{N}, \dots, E_{mn} = \frac{(\sum_{j=1}^{n} O_{mj})(\sum_{i=1}^{m} O_{in})}{N}$$

Chi-Square value is given by: $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

For a $m \times n$ contingency table (m rows and n columns), $\nu = (m-1)(n-1)$

Example 6 Test the hypothesis that flower colour is independent of leaf pattern using the following information:

Leaves ↓	Flower colour→		
	White	Red	Total
Flat	68	32	100
Curled	52	58	110
Total	120	90	210

Value of χ^2 for 2 and 1 degree of freedom at 5% level is 5.991 and 3.841 respectively.

Solution: Let H_0: Flower colour and leaf pattern are independent

On the basis of null hypothesis, expected (theoretical) frequencies are given as:

Leaves ↓	Flower colour→		
	White	Red	Total
Flat	$\frac{(100)(120)}{210} = 57.14$	$\frac{(100)(90)}{210} = 42.86$	100
Curled	$\frac{(110)(120)}{210} = 62.86$	$\frac{(110)(90)}{210} = 47.14$	110
Total	120	90	210

Degrees of freedom $(\nu) = (2-1)(2-1) = 1$

$$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(68-57.14)^2}{57.14} + \frac{(32-42.86)^2}{42.86} + \frac{(52-62.86)^2}{62.86} + \frac{(58-47.14)^2}{47.14} = 9.19$$

Table value of χ^2 for 1 degrees of freedom at 5% level = 3.841

Calculated value of χ^2 is greater than table value, $\therefore$ H_0 is rejected, i.e. flower colour and leaf pattern have some connection.

Example7 For the attributes A and B, in a contingency table $\begin{matrix} a & b \\ c & d \end{matrix}$

show that $\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$

Solution: The 2 × 2 contingency table for the attributes A and B is given below

Attribute	$B \rightarrow$		Total
A	a	b	$a+b$
$\downarrow$	c	d	$c+d$
Total	$a+c$	$b+d$	$N=a+b+c+d$

Here a, b, c, d are observed frequencies and under the assumption that the two attributes A and B are independent, expected frequencies are given as $\frac{(a+b)(a+c)}{N}$, $\frac{(a+b)(b+d)}{N}$, $\frac{(c+d)(a+c)}{N}$, $\frac{(c+d)(b+d)}{N}$ respectively.

$$\text{Now } \chi^2=\sum\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\frac{\left[a-\left(\frac{(a+b)(a+c)}{N}\right)\right]^2}{\frac{(a+b)(a+c)}{N}}+\cdots+\frac{\left[d-\left(\frac{(c+d)(b+d)}{N}\right)\right]^2}{\frac{(c+d)(b+d)}{N}}$$

$$=\frac{[a(a+b+c+d)-(a+b)(a+c)]^2}{(a+b+c+d)(a+b)(a+c)}+\cdots+\frac{[d(a+b+c+d)-(c+d)(b+d)]^2}{(a+b+c+d)(c+d)(b+d)}$$

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

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

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

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

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

Example8 A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by voting preferences to parties A, B and C and by gender (male or female). Results are shown in the contingency table below.

	Voting Preferences			Total
	Party A	Party B	Party C	
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

Do the men’s voting preferences differ significantly from the women’s preferences indicating a gender bias? Use a 0.05 level of significance.

Solution: Let H_0: Voting preferences and gender are independent.

	Voting Preferences			Total
	Party A	Party B	Party C	
Male	$\frac{(400)(450)}{1000} = 180$	$\frac{(400)(450)}{1000} = 180$	$\frac{(400)(100)}{1000} = 40$	400
Female	$\frac{(600)(450)}{1000} = 270$	$\frac{(600)(450)}{1000} = 270$	$\frac{(600)(100)}{1000} = 60$	600
Total	450	450	100	1000

On the basis of null hypothesis, expected frequencies are given in the above table.

Degrees of freedom $(\nu) = (2-1)(3-1) = 2$

$$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} + \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60}$$

$=$ 16.20, table value of χ^2 for 2 degrees of freedom at 5% level = 5.991

Calculated value of χ^2 is greater than table value, $\therefore$ H_0 is rejected, i.e. voting preferences and gender cannot be considered independent.

Exercise 7

1. Following tables shows number of male and female births in 800 families having four children:

Number of male births	0	1	2	3	4
Number of female births	4	3	2	1	0
Number of families	32	178	290	236	94

   Test if the data is consistent with the hypothesis that the binomial distribution holds and the probability of mail birth is same as that of a female birth

2. A random sample of 395 people was surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	Senior Secondary School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Test whether gender and education level independent at 5% level of significance? Value of χ^2 at 3 degrees of freedom is 7.815.

3. Following data shows number of good and bad parts produced by each of the three shifts in a factory.

	Good parts	Bad parts	Total
Day shift	960	40	1000
Evening shift	940	50	990
Night shift	950	45	995
Total	2850	135	2985

Test whether the production of bad parts is independent of the shift on which they were produced. Value of χ^2at 2 degrees of freedom is 5.991.

4. Out of a sample of 120 persons in a village, 76 persons were administered a new drug for controlling influenza and out of them, 24 persons were attacked by influenza. Out of those who were not administered the new drug, 12 persons were not affected by influenza. Use Chi-square test for finding whether the new drug is effective or not. Value of $\chi^2$for 1 degree of freedom is 3.84.
5. In a survey of 200 boys of which 75 are intelligent, 40 have educated fathers, while 85 of the unintelligent boys have uneducated fathers. Do these figures support the hypothesis that educated fathers have intelligent boys? Value of $\chi^2$for 1 degree of freedom is 3.84.
6. The following table shows the number of people having T.B. out of number of enquired people in different age- groups.

Age-group	Number of people enquired	T.B. cases
15-25	199	1
25-35	300	8
35-45	1128	38
45-55	1375	96
55-65	1089	105
65-75	625	56
75-85	155	12
Total	4871	316

Do these figures support the hypothesis that T. B. is equally spread in all age-groups?

Value of χ^2at 6 degrees of freedom is 12.59

7. Two hundred digits were chosen at random from a set of tables and their frequencies are given below:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	21	20	21	15

Test the hypothesis that the digits were distributed in an equal manner in the tables from which they are chosen. Value of χ^2at 9 degrees of freedom is 16.919.

Following data gives number of accidents in a city from years 2005 to 2014. Use χ^2 test of goodness of fit to prove the hypothesis that the number of accidents reported for each year from 2005 to 2014 does not differ significantly from an equal number of accidents in each year.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Number of Accidents	164	142	153	171	155	148	136	133	138	140

Answers

1. $\chi^2 = 54.43 \ \therefore$ probability of mail birth is not same as that of a female birth
2. $\chi^2 = 8.006$, which is greater than table value $\therefore$ gender and education levels seem dependent on each other.
3. $\chi^2 = 1.2\mathit{8}$, which is less than table value $\therefore$ production of bad parts is independent of the shifts
4. $\chi^2 = 18.97$, which is much greater than table value $\therefore$ null hypothesis is not supported, i.e. the new drug is definitely effective for controlling influenza.
5. $\chi^2 = 8.88$, $\therefore$ education of fathers has a significant effect on intelligence of boys.
6. $\chi^2 = 57.6$, which is much greater than table value $\therefore$ hypothesis is not supported.
7. $\chi^2 = 4.3$, which is less than table value $\therefore$ hypothesis is accepted.