Chapter 12

Chi-Square (χ^2) Distribution and its properties

Unit Structure:

12.1 Objectives

After going through this chapter students will be able to understand:

- The Chi-square distribution.
- The Chi-square test statistic.
- Uses of the Chi-square test.
- Pair of categorical variables can be summarized using contingency table.
- Perform a Chi-square goodness of fit test.
- The Chi-square test can compare an observed contingency table to an expected table and determine if the categorical variable are independent.
- YATE'S Correction for Contingency table.

12.2 Introduction:

Chi-square (χ^2) test is a nonparametric statistical analyzing method often used in experimental work where the data consist in frequencies or 'counts' – for example the number of boys and girls in a class having their tonsils out – as distinct from quantitative data obtained from measurement of continuous variables such as temperature, height, and so on. The most common use of the test is to assess the probability of association or independence of facts.

A common problem in applied machine learning is determined whether input features are relevant to the outcome to be predicated. This is the problem of feature selection.

In the case of classification problems where input variables are also categorical, we can use statistical tests to determine whether the output variable is dependent or independent of the input variables. If independent then the input variable is a candidate for a feature that may be irrelevant to the problem and removed from the dataset.

The Pearson's Chi-square statistical hypothesis is an example of a test for independence between categorical variables.

In this chapter, you will learn the Chi-square statistical hypothesis test for quantifying the independence of pairs of categorical variables.

12.3 Chi-Square distribution:

We have been discussing the distribution of mean obtained from all possible samples or a large number of samples drawn from a normal population, distribution with mean μ and variance $\frac{\sigma^2}{n}$. Now we are interested in knowing the distribution of sample variances s^2 of these samples. Consider a random sample $X_1, X_2, \ldots\ldots., X_n$ of size n. Let the observations of this sample be denoted by $x_1, x_2, \ldots.., x_n$. We know that the variance,

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \text{ for } i = 1, 2, \ldots\ldots.., n$$

Or $\sum_i (x_i - \bar{x})^2 = (n-1)s^2 = ks^2$ where $k = (n-1)$.

A quantity $\frac{ks^2}{\sigma^2}$, which is a pure number is defined as χ^2_k. Now we will give the distribution of the random variable χ^2_k, which was first discovered by Helmert in 1876 and later independently given by Karl Pearson in 1900. Another way to understand Chi-square is : if $X_1, X_2, \ldots\ldots., X_n$ are n independent normal variates with mean zero and variance unity, the sum of squares of these variates is distributed as Chi-square with n degree of freedom. The Chi-square distribution was discovered mainly as a measure of goodness of fit in case of frequency distribution, i.e. whether the observed frequencies follow a postulated distribution or not. The probability density function of χ^2- variate is,

$$f_k\left(\chi^2\right) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}\left(\chi^2\right)^{\frac{1}{2}k-1} e^{\frac{-1}{2}\chi^2}$$

12.4 Properties of Chi-Square Distribution:

The χ^2 distribution follow the following properties:

1. The whole χ^2 distribution curve lies in the first quadrant since the range of χ^2 is from 0 to ∞.
2. The χ^2 distribution has only one parameter k, the degree of freedom for χ^2. Thus, the shape of the probability density curve mainly depends on the parameter k.
3. χ^2 distribution curve is highly positive skewed.
4. It is an unimodal curve and its Mode is at the point $\chi^2 = (k - 1)$.
5. The shape of the curve varies immensely especially when k is small. For k =1 and k = 2, it is just like a hyperbola.
6. χ^2 distribution is completely defined by one parameter 'k', which is known as the degree of freedom for χ^2 distribution.
7. The constants for χ^2 distributions are as follows:
   Mean = $\mu = k$
   Variance =$\sigma^2 = 2k$
   Skewness = $\alpha_1 = 2\left(\frac{2}{k}\right)^{\frac{1}{2}}$
8. The movement generating function for χ^2 distribution is
   $$\emptyset_{\chi^2}(t) = (1 - 2t)^{\frac{-k}{2}}.$$
9. r^{th} raw moment of χ^2 distribution is $\mu'_r = \frac{2\Gamma\left(\frac{k}{2}+r\right)}{\frac{\Gamma k}{2}}$.
10. For large degrees of freedom say $k \geq 100$, the variable is distributed normally with mean 0 and variance 1.

12.5 Test of Goodness of Fit:

Generally the population study has been taken to follow a known distribution such as normal, binomial or Poisson distribution. To assume that the population is distributed normally is a common practice and hence we explain the test of goodness of fit of normal population first.

This is very powerful test for testing difference between observed data and theoretical expectation. The test is given by Karl-Pearsons and is known as χ^2 test of goodness of fit.

The test statistic used for χ^2 distribution is based on two types of frequencies namely observed frequencies denoted by O_i and expected frequencies denoted by E_i. The larger the deviation from the null hypothesis, the larger the difference between observed and expected frequencies then Karl-Pearson's χ^2 is given by

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

Follow χ^2 distribution with $(k - p - 1)$ degree of freedom (d.f).

- If the calculated value of χ^2 is greater than the table value of χ^2 for $(k - p - 1)$ d.f. and level of significance α, reject H_0. Rejection of H_0 means that the postulated theoretical distribution is not fit to the observed data, or in other words the data do not support the assertion about the theoretical distribution.

Procedure for test significance and goodness of fit:

Set up a null hypothesis and calculate $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

1. Find the degree of freedom (d.f.) and read the corresponding value of χ^2 at a prescribed significance level from table.
2. From of χ^2 table we can also find the probability P corresponding to the calculated values of χ^2 for the given d.f.
3. If $P < 0.05$ the observed value of χ^2 is significant at 5% level of significance.
4. If $P > 0.05$ it is good fit and the value is not significant.

12.6 Contingency table:

The data are often based on counting of objects of units. These numbers fall in various categories of attributes in a two-way classification and are very well displayed systematically in a table know as contingency table. We can express a contingency table as a rectangular array of order $(m \times n)$, having mn cells, where m denotes the number of rows which are equal to the number of categories of an attribute or criterion X and n denotes the number of columns equal to the number of categories of an attribute or criterion Y.

Attribute X	Attribute Y	Total
	Y_1 Y_2..............Y_n	
X_1 X_2. . . . X_m		
Total		

2×2 **Contingency table:**

When a contingency table is of order 2×2, test of independence of factors can be performed in the same manner as for $(m \times n)$ contingency table. But in this situation the value of χ^2 can also be calculated directly from the observed frequencies. It is nothing but a short-cut method to obtaining the calculated value of χ^2. Suppose the contingency table of order 2×2 for two factor X and Y is as presented below.

Factor X	Factor Y		Total
	Y_1	Y_2	
X_1	a	b	(a + b)
X_2.	c	d	(c + d)
Total	(a +c)	(b+d)	a + b + c + d = n

12.7 Test of Independence of Factors

It is apparent now, that a contingency table is a rectangular array having rows and columns associated with different factors. The hypothesis that one factor is independent of the other or not, i.e. H_0 : Two factors are independent of each other.

H_1 : Two factors are dependent of each other.

The test statistic used for χ^2 distribution is based on two types of frequencies namely observed frequencies denoted by O_i and expected frequencies denoted by E_i. The larger the deviation from the null hypothesis, the larger the difference between observed and expected is.

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

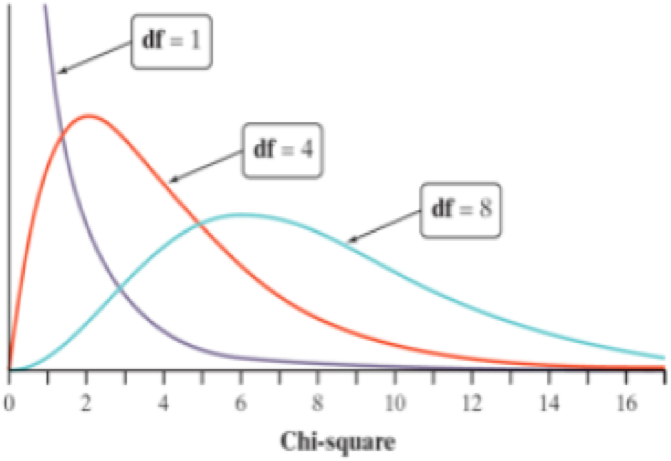
Squaring the differences makes them all positive. Each difference is divided by the expected number, and these standardized ratios are summed: the more differences between what you would expect and what you get the bigger the number.

Degrees of Freedom: A critical factor in using the chi-square test is the "degrees of freedom", which is essentially the number of independent random variables involved. For example, you do a cross and see 290 purple flowers and 110 white flowers in the offspring.

- Degrees of freedom is simply the number of classes of offspring minus 1.
- For our example, there are 2 classes of offspring: purple and white. Thus, degrees of freedom (d.f.) = 2 -1 = 1.

Number of degrees of freedom = (number of rows -1) (number of columns – 1)

i.e. $d.f = (r - 1)(c - 1)$



- The image above shows that the distribution of the Chi-square statistic starts at zero and can only have positive values.
- The shape of the distribution is much different than the t or z statistic and is skewed to the right.
- The shape of the distribution changes as the degree of freedom increses.

Critical Chi-Square: Critical values for chi-square are found on tables, sorted by degrees of freedom and probability levels. Be sure to use p = 0.05.

If your calculated chi-square value is greater than the critical value from the table, you "reject the null hypothesis". If your chi-square value is less than the critical value, you "fail to reject" the null hypothesis (that is, you accept that your genetic theory about the expected ratio is correct).

Critical values of the χ^2 Distribution table

d.f.	α								
	0.995	**0.975**	**0.9**	**0.5**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	**0.000**	**0.000**	**0.016**	**0.455**	**2.706**	**3.841**	**5.024**	**6.635**	**7.879**
2	**0.010**	**0.051**	**0.211**	**1.386**	**4.605**	**5.991**	**7.378**	**9.210**	**10.597**
3	**0.072**	**0.216**	**0.584**	**2.366**	**6.251**	**7.815**	**9.348**	**11.345**	**12.838**
4	**0.207**	**0.484**	**1.064**	**3.357**	**7.779**	**9.488**	**11.143**	**13.277**	**14.860**
5	**0.412**	**0.831**	**1.610**	**4.351**	**9.236**	**11.070**	**12.832**	**15.086**	**16.750**
6	**0.676**	**1.237**	**2.204**	**5.348**	**10.645**	**12.592**	**14.449**	**16.812**	**18.548**
7	**0.989**	**1.690**	**2.833**	**6.346**	**12.017**	**14.067**	**16.013**	**18.475**	**20.278**
8	**1.344**	**2.180**	**3.490**	**7.344**	**13.362**	**15.507**	**17.535**	**20.090**	**21.955**
9	**1.735**	**2.700**	**4.168**	**8.343**	**14.684**	**16.919**	**19.023**	**21.666**	**23.589**
10	**2.156**	**3.247**	**4.865**	**9.342**	**15.987**	**18.307**	**20.483**	**23.209**	**25.188**
11	**2.603**	**3.816**	**5.578**	**10.341**	**17.275**	**19.675**	**21.920**	**24.725**	**26.757**
12	**3.074**	**4.404**	**6.304**	**11.340**	**18.549**	**21.026**	**23.337**	**26.217**	**28.300**
13	**3.565**	**5.009**	**7.042**	**12.340**	**19.812**	**22.362**	**24.736**	**27.688**	**29.819**
14	**4.075**	**5.629**	**7.790**	**13.339**	**21.064**	**23.685**	**26.119**	**29.141**	**31.319**
15	**4.601**	**6.262**	**8.547**	**14.339**	**22.307**	**24.996**	**27.488**	**30.578**	**32.801**

Note:

1. χ^2 test is non-parametric test.
   Use Nonparametric Tests:

- Used when either the dependent or independent variable is ordinal.

- Used when the sample size is small.
- Used when underlying population is not normal.

2. The value of χ^2 test statistic can never be negative.

Example 1: In experiments on pea breading the following frequencies of seeds were obtained:

Round and yellow	Wrinkled and yellow	Round and green	Wrinkled and green	Total
315	101	108	32	556

Theory predicate that the frequencies should be in proportions 9:3:3:1. Examine the correspondence between theory and experiment.

Solution: First select null hypothesis $H_0 =$ The correspondence between theory and experiment.

Theory predicate that the frequencies should be in proportions 9:3:3:1.

Total value = 9 + 3 + 3 + 1= 16

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
315	$\frac{9}{16}\times556 = 313$	4	0.0128
101	$\frac{3}{16}\times556 = 104$	9	0.0865
108	$\frac{3}{16}\times556 = 104$	16	0.1538
32	$\frac{1}{16}\times556 = 35$	9	0.2571
		Total	0.5102

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 0.5102$$

Number of degrees of freedom $= (k - 1) = (4 - 1) = 3$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 7.815.

$\Rightarrow \chi^2_{cal} < \chi^2_{tab}$

Therefore, H_0 is accepted.

Hence there is a very high degree of agreement between theory and experiment.

Example 2: A set of five similar coins is tossed 320 times and the result is

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution.

Solution: First select null hypothesis, H_0 = The data follow a binomial distribution.

P: probability of getting a head = $\frac{1}{2}$

q: probability of not getting a head = $\frac{1}{2}$

Here for expected frequencies,

P(zero head) =$^5C_0\, p^0 q^5 \times 320 = \frac{1}{32} \times 320 = 10$

P(one head) =$^5C_1\, p^1 q^4 \times 320 = \frac{5}{32} \times 320 = 50$

P(Two head) =$^5C_2\, p^2 q^3 \times 320 = \frac{10}{32} \times 320 = 100$

P(Three head) =$^5C_3\, p^3 q^2 \times 320 = \frac{10}{32} \times 320 = 100$

P(Four head) =$^5C_4\, p^4 q^1 \times 320 = \frac{5}{32} \times 320 = 50$

P(Five head) =$^5C_5\, p^5 q^0 \times 320 = \frac{1}{32} \times 320 = 10$

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
6	10	16	1.6
27	50	529	10.58
72	100	784	7.84
112	100	144	1.44
71	50	441	8.82
32	10	484	48.4
		Total	78.68

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 78.68$$

Number of degrees of freedom $= (k - 1) = (6 - 1) = 5$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 11.07.

$\Rightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence the data follow the binomial distribution is rejected.

Example 3: Fit a Poisson distribution to the following data and test for its goodness of fit at level of significance 0.05.

X	0	1	2	3	4
F	419	352	154	56	19

Solution: First select null hypothesis, H_0 = The data follow a Poisson distribution.

For Poisson distribution we need to find mean of the data.

X	F	fx
0	419	0
1	352	352
2	154	308
3	56	168
4	19	76
Total	1000	904

$$m = \frac{\Sigma fx}{\Sigma f} = \frac{904}{1000} = 0.904$$

$$\therefore e^{-0.904} = 0.4049$$

Here for expected frequencies are $\frac{e^{-0.904} \times 0.904^{x}}{x!} \times 1000.$

P(x = 0)= $\frac{e^{-0.904}\times 0.904^{x}}{x!}\times 1000\ =\ 404.9$

P(x = 1)= $\frac{e^{-0.904}\times 0.904^{x}}{x!}\times 1000\ =\ 366$

P(x = 2)= $\frac{e^{-0.904}\times 0.904^{x}}{x!}\times 1000\ =\ 165.4$

P(x = 3)= $\frac{e^{-0.904}\times 0.904^{x}}{x!}\times 1000\ =\ 49.8$

P(x = 4)= $\frac{e^{-0.904}\times 0.904^{x}}{x!}\times 1000\ =\ 11.3$

In order that the total observed and expected frequencies may agree, we take the first and last theoretical frequencies as 406.2 and 12.6 instead of 404.9 and 11.3 as shown in table.

Therefore expected frequencies distribution

X	0	1	2	3	4	Total
F	404.9	366	165.4	49.8	11.3	997.4
Instead	406.2				12.6	1000

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
419	406.2	163.84	0.403
352	366	196	0.536
154	165.4	129.96	0.786
56	49.8	38.44	0.772
19	12.6	40.96	3.251
		Total	5.748

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 5.748$$

Since the mean of the theoretical distribution has been estimated from the given data and the totals have been made to agree, there are two constrains so that the number of degree of freedom

Number of degrees of freedom $= (k\ -\ 2) = 5\ -\ 2\ =\ 3$

χ^2 table value for 3 d.f at $\alpha\ =\ 0.05$ = 7.815.

$\Rightarrow \chi^2_{cal} < \chi^2_{tab}$

Therefore, H_0 is accepted.

Hence, the Poisson distribution can be fitted to the data.

Example 4: A company director is concerned that his company's share may be unevenly distributed throughout the country, in a survey in which sample of 200 customers are selected from four zones and are tabulated as under:

	Zone				
	I	II	III	IV	
Purchase the brand	80	110	90	100	380
Not purchase the brand	120	90	110	100	420
	200	200	200	200	800

At $p = 0.05$, use χ^2 to determine whether the company share is same across the four zones.

Solution: First select null hypothesis H_0 and Alternative hypothesis H_1.

H_0 : The company share is same across four zones.

H_1 : The company share is not same across four zones.

At $\alpha = 0.05$.

Now calculate expected frequencies for given table:

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{380\times200}{800} = 95$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{380\times200}{800} = 95$$

$$E_{13} = \frac{R_1 \times C_3}{N} = \frac{380\times200}{800} = 95$$

$$E_{14} = \frac{R_1 \times C_4}{N} = \frac{380\times200}{800} = 95$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{420\times200}{800} = 105$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{420\times200}{800} = 105$$

$$E_{23} = \frac{R_2 \times C_3}{N} = \frac{420\times200}{800} = 105$$

$$E_{24} = \frac{R_2 \times C_4}{N} = \frac{420\times200}{800} = 105$$

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
80	95	225	2.368
110	95	225	2.368
90	95	25	0.263
100	95	25	0.263
120	105	225	2.143
90	105	225	2.143
110	105	25	0.238
100	105	25	0.238
		Total	10.024

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 10.024$$

Number of degrees of freedom = $(r - 1)(c - 1) = (2 - 1)(4 - 1) = 3$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 7.815.

$\Rightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence the company share is not same across four zones.

Example 5: From the following table showing the number of plants having certain characters, test the hypothesis that the flower colour is independent of flatness of leaf at the 0.1 level of significance.

	Flat leaves	Curled leaves	Total
White Flowers	99	36	135
Red Flowers	20	5	25
Total	119	41	160

Solution: First select null hypothesis H_0 and Alternative hypothesis H_1.

H_0 : The flower colour is independent of flatness of leaf.

H_1 : The flower colour is dependent of flatness of leaf.

At $\alpha = 0.1$.

Now calculate expected frequencies for given table:

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{135 \times 119}{160} = 100$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{135 \times 41}{160} = 35$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{25 \times 119}{160} = 19$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{25 \times 41}{160} = 6$$

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
99	100	1	0.01
36	35	1	0.0286
20	19	1	0.0526
5	6	1	0.1667
		Total	0.2579

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 0.2579$$

Number of degrees of freedom = $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

χ^2 table value for 1 d.f at $\alpha = 0.1$ is 0.0185

$\Rightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence The flower colour is dependent of flatness of leaf.

12.8 YATES Correction:

We know that the χ^2- distribution is a continuous distribution. It has been proved that if any of the cell frequency in contingency table of order 2×2 is less than 5, the continuity of χ^2- distribution curve is not maintained. So to remove this discrepancy, Yate's suggested a correction which is extensively used. He suggested that add 0.5 in the frequency which is less than 5, and subtract and add 0.5 to the remaining cell frequencies in such a way that the marginal totals remain the same. Then calculate the value of χ^2 by formula

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

Using adjusted contingency table.

Instead of adjusting the contingency table of order 2×2, the above formula has been amended and this takes care of the correction. The value of χ^2 under correction can directly be calculated by the formula,

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

Here $|ad - bc|$ means that we consider only the absolute value of the difference. Moreover, if one does not use the formula given specifically for the 2×2 contingency table but follows the general procedure, under Yate's correction, the formula for the χ^2 is,

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(|O_{ij}-E_{ij}|-\frac{1}{2}\right)^2}{E_{ij}}$$

It is worthwhile to point out that all three approaches yield the same value of the χ^2 because they are fundamentally the same.

Example 6: Use Yate's correction and test whether A and B are independent. Observed frequencies are as under:

	A	Not A	Total
B	45	55	100
Not B	60	40	100
Total	105	95	200

Solution:

H_0: Two attribute A and B are independent.

To use χ^2 test with Yate's correction, we have

$$n = 200, \ a = 45, \ b = 55, \ c = 60, \ d = 40$$

The χ^2 test statistic is

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

$$\chi^2 = \frac{200\left(|45\times40-55\times60|-\frac{200}{2}\right)^2}{(105)(95)(100)(100)}$$

$$\chi^2 = \frac{200\left(|1800-3300|-\frac{200}{2}\right)^2}{(105)(95)(100)(100)}$$

$$\chi^2 = \frac{200(1500-100)^2}{(105)(95)(100)(100)}$$

$$\chi^2 = \frac{200\times1400\times1400}{(105)(95)(100)(100)} = 3.9298$$

Number of degrees of freedom = $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

χ^2 table value for 1 d.f at $\alpha = 0.05$ is 3.84

$\Rightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence A and B are dependent.

Example 7: the number of licensor companies classified by the proportion of foreign profits derived from license agreements were as follows:

Proportion of profits	Licensor type		Total
	Dominant product	Diversified	
Less than 5% 5% or more	1 7	6 6	7 13
Total	8	12	20

Solution : Let H_0: Proportion of profit and licensor types are independent.

To use χ^2 test with Yate's correction, we have

$$n = 20, \; a = 1, \; b = 6, \; c = 7, \; d = 6$$

The χ^2 test statistic is

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

$$\chi^2 = \frac{20\left(|1\times6-6\times7| - \frac{20}{2}\right)^2}{(1+7)(6+6)(1+6)(7+6)}$$

$$\chi^2 = \frac{20(36-10)^2}{8\times12\times7\times13} = \frac{13520}{8736} = 1.55$$

Number of degrees of freedom = $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

χ^2 table value for 1 d.f at $\alpha = 0.05$ is 3.841

$\Rightarrow \chi^2_{cal} < \chi^2_{tab}$

Therefore, H_0 is accepted.

Hence, Proportion of profit and licensor types are independent .

12.9 Lets sum up

In this chapter we have learnt the following:

- Chi-square Distribution and its properties.
- Goodness of fit for Chi-square distribution.
- Uses of the Chi-square test.
- Pair of categorical variables can be summarized using contingency table..
- The Chi-square test can compare an observed contingency table to an expected table and determine if the categorical variable are independent.
- YATE'S Correction for Contingency table.

12.10 Unit End exercise

1. The following table is given

Eye colour in fathers	Eye colour in sons		Total
	Brown	Black	
Brown	230	148	378

Black	251	471	622
Total	381	619	1000

Test whether the colour of the son's eyes is associated with that of the fathers.

2. Fit a Poisson distribution to the following data and test the goodness of fit.

x	0	1	2	3	4	5	6
f	275	72	30	7	5	2	1

3. A large city fire department calculates that for any given instance, during any given 8 hrs shift, there is a 30% chance of recurring at least one fire alarm. Here is a random sampling of 60 days:

No. of shift during which alarms were received	0	1	2	3
No. of days	16	27	11	6

At 5% level of significance verify that binomial distribution fits the data.

4. Test the hypothesis that the following observations follow a Poisson distribution with mean 4. Use 5% as level of significance.

No.of call per hrs.	0	1	2	3	4	5
No. of hrs.	20	57	98	85	78	62

5. Genetic theory states that children having one parent of blood type A and other blood type B will always be one of three types A, AB, B and that proportions of these types will on average be 1:2:1. A report states that out of 300 children having one A parent and one B parent, 30% were found to be type A , 45% of type AB and remaining of type B. Test the hypothesis by Chi-square test.
6. Fit a Binomial distribution to the data:

x	0	1	2	3	4	5
f	38	144	342	287	164	25

And test for goodness of fit at the level of significance 0.05.

7. The table shows the relation between the performance in mathematics and IT, using 0.01 significance level.

Mathematics Marks	IT Marks			Total
	High	Medium	Low	
High	56	71	12	139
Medium	47	163	38	248
Low	14	42	85	141
Total	117	276	135	528

8. In an experiment on immunization of human from COVID-19 the following results were obtained.

	Died	Unaffected
Inoculated	12	26
Not inoculated	16	6

Examine the effect of vaccine in controlling susceptibility to COVID-19.

9. A die is thrown 120 times with the following results:

Face	1	2	3	4	5	6
Frequency	16	30	22	18	14	20

Test the hypothesis that the die is unbiased at level of 5% significance.

10. In a survey of 200 boys, of which 75 were intelligent, 40 had skilled fathers; while 85 of the unintelligent boys had unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys? Use Chi-square test at 5% level of signification.
11. Among 64 offspring's of a certain cross between guinea pigs, 34 were red, 10 were black and 20 were white. According to the genetic model, these numbers should be in the ratio 9:3:4, are the data consistent with the model at the 5% level?
12. Opinion about promotions, to be dependent on published work by persons interested in teaching or research was taken and displayed as below

Interest	Promotion dependent On published work		Total
	Agree	Disagree	
Teaching	90	10	100
Research	70	30	100

Total	160	40	200

Examine whether the promotion dependent on published work and interest.

12.11 Reference

Fundamentals of mathematical statistics by S.C. Gupta and V.K. Kapoor

A Guide to Chi-Squared Testing Priscilla E. Greenwood, Michael S. Nikulin