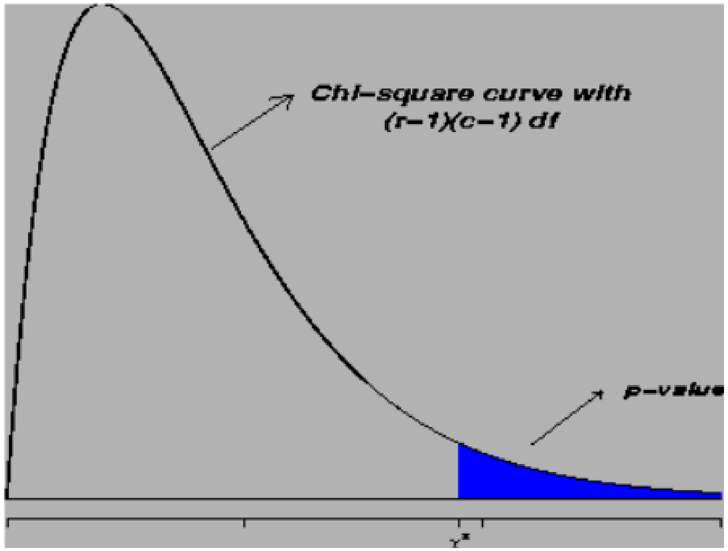
Introduction

- The Chi-square test is one of the most commonly used non-parametric test, in which the sampling distribution of the test statistic is a ***chi-square distribution***, when the null hypothesis is true.
- It was introduced by ***Karl Pearson*** as a test of association. The Greek Letter $\chi 2$ is used to denote this test.
- It can be applied when there are few or no assumptions about the population parameter.
- It can be applied on categorical data or qualitative data using a contingency table.
- Used to evaluate ***unpaired/unrelated samples and proportions.***

Chi-squared distribution

- The distribution of the chi-square statistic is called the chi-square distribution.
- The **chi-squared distribution** with *k* degrees of freedom is the distribution of a sum of the squares of *k* independent standard normal random variables. It is determined by the *degrees of freedom.*
- The simplest chi-squared distribution is the square of a standard normal distribution.
- The chi-squared distribution is used primarily in hypothesis testing.



Contingency table

- A **contingency table** is a type of table in a matrix format that displays the frequency distribution of the variables.
- They provide a basic picture of the interrelation between two variables and can help find interactions between them.

	Column 1	Column 2	Totals
Row 1	A	B	R1
Row 2	C	D	R2
Totals	C1	C2	N

- The chi-square statistic compares the observed count in each table cell to the count which would be expected ***under the assumption of no association between the row and column classifications.***

Degrees of freedom

- The number of independent pieces of information which are free to vary, that go into the estimate of a parameter is called the degrees of freedom.
- In general, the degrees of freedom of an estimate of a parameter is equal to ***the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself*** (i.e. the sample variance has *N*-1 degrees of freedom, since it is computed from *N* random scores minus the only 1 parameter estimated as intermediate step, which is the sample mean).
- The number of degrees of freedom for 'n' observations is 'n-k' and is usually denoted by 'ν', where 'k' is the number of independent linear constraints imposed upon them. It is the only parameter of the chi-square distribution.
- The degrees of freedom for a chi squared contingency table can be calculated as:

$$\nu = (\text{Number of rows} - 1) * (\text{Number of columns} - 1)$$

SURESH GYAN VIHAR UNIVERSITY
Accredited by NAAC with 'A' Grade

Chi Square formula

- The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.
- The value of χ 2 is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \ldots\ldots + \frac{(O_n - E_n)^2}{E_n}$$

Where, O_1, O_2, O_3….On are the observed frequencies and E_1, E_2, E_3…E_n are the corresponding expected or theoretical frequencies.

The observed frequencies are the frequencies obtained from the observation, which are sample frequencies.
The expected frequencies are the calculated frequencies.

Characteristics of Chi-Square test

1. It is often regarded as a ***non-parametric test*** where no parameters regarding the rigidity of populations are required, such as mean and SD.
2. It is based on ***frequencies***.
3. It encompasses the ***additive property*** of differences between observed and expected frequencies.
4. It tests the hypothesis about the ***independence of attributes***.
5. It is preferred in analyzing complex contingency tables.

Steps in solving problems related to Chi-Square test

STEP 1

- Calculate the expected frequencies $E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$

STEP 2

- Take the difference between the observed and expected frequencies and obtain the squares of these differences $(O\text{-}E)^2$

STEP 3

- Divide the values obtained in Step 2 by the respective expected frequency, E and add all the values to get the value according to the formula given by:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

SURESH GYAN VIHAR UNIVERSITY
Accredited by NAAC with 'A' Grade

Conditions for applying Chi-Square test

1. The data used in Chi-Square test must be ***quantitative*** and in the form of ***frequencies***, which must be ***absolute*** and not in relative terms.
2. The total number of observations collected for this test must be ***large*** (at least 10) and should be done on a ***random*** basis.
3. Each of the observations which make up the sample of this test must be ***independent*** of each other.
4. The expected frequency of any item or cell must not be ***less than 5***; the frequencies of adjacent items or cells should be polled together in order to make it more than 5.
5. This test is used only for ***drawing inferences*** through test of the hypothesis, so it ***cannot be used for estimation*** of parameter value.

Practical applications of Chi-Square test

- The applications of Chi-Square test include testing:

1. The significance of ***sample & population variances*** **[σ^2 s & σ^2 p]**
2. The ***goodness of fit*** of a theoretical distribution: Testing for goodness of fit determines if an observed frequency distribution fits/matches a theoretical frequency distribution (**Binomial distribution, Poisson distribution or Normal distribution)**. These test results are helpful to know whether the samples are drawn from identical distributions or not. **When the calculated value of χ^2 is less than the table value at certain level of significance, the fit is considered to be good one and if the calculated value is greater than the table value, the fit is not considered to be good.**

Table/Critical values of χ^2

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

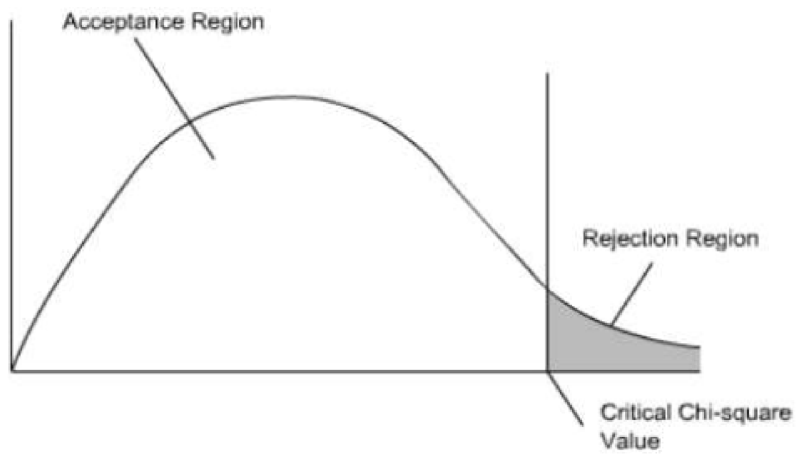
3. The ***independence*** in a contingency table:
   - Testing independence determines whether two or more observations across two populations are dependent on each other.
   - If the **calculated value is less than the table value** at certain level of significance for a given degree of freedom, then it is concluded that null hypothesis is true, which means that two attributes are independent and hence not associated.
   - If **calculated value is greater than the table value**, then the null hypothesis is rejected, which means that two attributes are dependent.
4. The chi-square test can be used to test the strength of the association between exposure and disease in a ***cohort study, an unmatched case-control study, or a cross-sectional study***.

Interpretation of Chi-Square values

- The χ 2 statistic is calculated under the ***assumption of no association***. „
- **Large value of χ 2 statistic** $\Rightarrow$ ***Small probability*** of occurring by chance alone ($p < 0.05$) $\Rightarrow$ Conclude that ***association exists*** between disease and exposure. „ **(Null hypothesis rejected)**
- **Small value of χ 2 statistic** $\Rightarrow$ ***Large probability*** of occurring by chance alone ($p > 0.05$) $\Rightarrow$ Conclude that ***no association exists*** between disease and exposure. **(Null hypothesis accepted)**

Interpretation of Chi-Square values

- The left hand side indicates the degrees of freedom. If the calculated value of $\chi 2$ falls in the acceptance region, the null hypothesis 'Ho' is accepted and vice-versa.



Limitations of the Chi-Square Test

1. The chi-square test does ***not give us much information about the strength of the relationship***. It only conveys the existence or nonexistence of the relationships between the variables investigated.
2. The chi-square test is ***sensitive to sample size***. This may make a weak relationship statistically significant if the sample is large enough. Therefore, chi-square should be used together with measures of association like ***lambda, Cramer's V or gamma*** to guide in deciding whether a relationship is important and worth pursuing.
3. The chi-square test is also ***sensitive to small expected frequencies***. It can be used only when not more than **20%** of the cells have an ***expected frequency of less than 5.***
4. Cannot be used when samples are ***related or matched***.

EXAMPLES:

Estrogen supplementation to delay or prevent the onset of Alzheimer's disease in postmenopausal women.

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147	9	156
	No	810	158	968
		957	167	1,124

The null hypothesis (H_0): Estrogen supplementation in postmenopausal women is unrelated to Alzheimer's onset.

The alternate hypothesis(H_A): Estrogen supplementation in postmenopausal women delays/prevents Alzheimer's onset.

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147	9	156
	No	810	158	968
		957	167	1,124

Of the women who did not receive estrogen supplementation, 16.3% (158/968) showed signs of Alzheimer's disease onset during the five-year period; whereas, of the women who did receive estrogen supplementation, only 5.8% (9/156) showed signs of disease onset.

- **Next step: To calculate expected cell frequencies**

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147	9	156
	No	810	158	968
		957	167	1,124

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	$E_a = \frac{156\text{x}957}{1124} = 132.82$	$E_b = \frac{156\text{x}167}{1124} = 23.18$	156
	No	$E_c = \frac{968\text{x}957}{1124} = 824.18$	$E_d = \frac{968\text{x}167}{1124} = 143.82$	968
		957	167	1,124

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147 132.82	9 23.18	156
	No	810 824.18	158 143.82	968
		957	167	1,124

$$\chi^2 = \sum \frac{(|O - E| - .5)^2}{E}$$

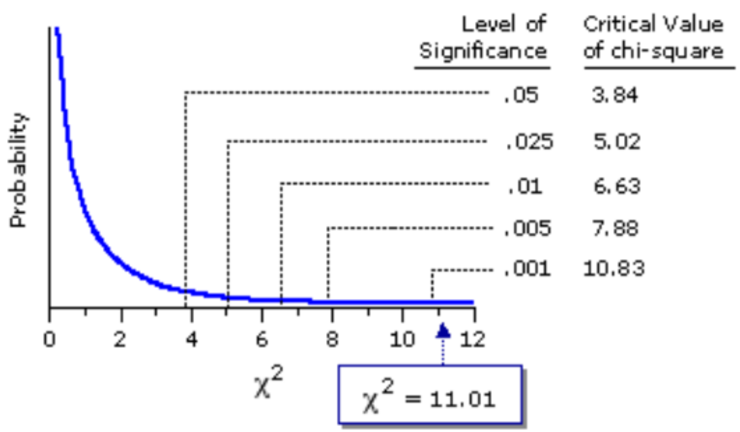
		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	$\frac{(\lvert 147-132.82 \rvert -.5)^2}{132.82}$ $= 1.41$	$\frac{(\lvert 9-23.18 \rvert -.5)^2}{23.18}$ $= 8.07$	
	No	$\frac{(\lvert 810-824.18 \rvert -.5)^2}{824.18}$ $= 0.23$	$\frac{(\lvert 158-143.82 \rvert -.5)^2}{143.82}$ $= 1.3$	
				sum: $\chi^2 = 11.01$

The next step is to refer calculated value of chi-square to the appropriate sampling distribution, which is defined by the applicable number of degrees of freedom.

- For this example, there are 2 rows and 2 columns. Hence,

df = (2—1)(2—1) = 1

Sampling Distribution of Chi-Square for df=1



- The calculated value of $\chi2$ =11.01 exceeds the value of chi-square (10.83) required for significance at the 0.001 level.
- Hence we can say that the observed result is significant *beyond* the 0.001 level.
- Thus, the null hypothesis can be rejected with a high degree of confidence.