Chi-square test (χ^2 -test)

The **chi-square test** is a statistical method used to compare observed data with expected data based on a specific hypothesis. It helps determine whether there is a significant difference between the expected and observed frequencies in categorical data.

Assumptions of the Chi-Square Test

- The data should be in the form of frequencies or counts of categories.
- Observations must be independent.
- The expected frequency in each category should be at least 5 for accurate results.
- The total sample size should be sufficiently large.

Chi-Square Test for Independence of Attributes:

The Chi-Square Test for Independence is a statistical test used to determine whether two categorical variables are associated or independent of each other. It compares observed frequencies (actual data) in a contingency table to the frequencies expected if the variables were independent.

Conditions for Using the Test

1. **Two Categorical Variables**: The test is used when your data involves two variables that can be categorized (e.g., gender, preferences, education level). Examples:
    a) "Does smoking status (smoker/non-smoker) depend on gender (male/female)?"
    b) "Is product preference (A/B/C) related to age group (child/teen/adult)?"
2. **Data Represented in Counts or Frequencies**: The data must represent counts or frequencies, not percentages, ratios, or continuous data. Example: The number of males and females preferring different products.
3. **Random Sampling**: The sample should be randomly selected from the population.
4. **Independence of Observations**: Each observation should be independent of others (e.g., one person cannot belong to multiple categories).
5. **Sufficient Sample Size**: The expected frequency in each cell of the contingency table should generally be at least 5 for the test to be reliable.

Examples of When to Use

1. **Market Research**: Determining if age group (young, middle-aged, senior) is associated with the choice of a streaming service (Netflix, Hulu, Amazon).
2. **Medical Studies**: Testing whether smoking status (smoker/non-smoker) is related to the presence of a specific disease (yes/no).
3. **Education Studies**: Evaluating if study habits (daily/weekly/monthly) are associated with performance levels (high/medium/low).
4. **Social Science Research**: Investigating if marital status (single/married/divorced) is related to political affiliation (party A/party B/independent).

Steps for Conducting the Test

1. Hypothesis Set up:

H_0: There is no relationship between two attributes or two attributes are independent.

H_1: There is relationship between two attributes or two attributes are dependent.

2. Test Statistics: Under H_0, the χ^2 -statistic is

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E}\right)$$

where, O = observed frequencies

E = expected frequencies = $\frac{\text{RT}*\text{CT}}{\text{N}}$

N = total sample size = grand total

For example, 2 x 2 contingency table:

	B_1	B_2	Row total (RT)
A_1	a	b	a+b
A_2	c	d	c+d
Column total (CT)	a+c	b+d	N= a+b+c+d

The expected frequencies for each cell can be obtained as follows:

$E\ (a) = \frac{\text{RT}*\text{CT}}{\text{N}} = \frac{(\text{a+b})(a+c)}{\text{N}}$, $\quad E\ (b) = \frac{(\text{a+b})(b+d)}{\text{N}}$

$E\ (c) = \frac{(\text{c+d})(a+c)}{\text{N}}$ and $E\ (d) = \frac{(\text{c+d}))(b+d)}{\text{N}}$

For 2 x 2 contingency table χ^2 can be computed using following formula:

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

3. Level of Significance: α

4. Degree of Freedom = (r-1) (c-1)

5. Critical Value: We have to determine the tabulated value of χ^2 at α% level of significance for (r-1) (c-1) degree of freedom from χ^2 table

6. Decision:

- If $\chi^2_{cal} \leq \chi^2_{tab}$, we do not reject H_0
- If $\chi^2_{cal} > \chi^2_{tab}$, we reject H_0

Questions

1. Four hundred employees of a certain factory were classified according to the gender and their level of satisfaction.

Gender	Level of satisfaction		
	Unsatisfied	Satisfied	Highly satisfied
Male	60	80	70
Female	40	60	90

Is there any relationship between gender and level of satisfaction? (χ^2_{cal} = 8.377, reject H_0)

2. A sample of 300 students of undergraduate and 300 of Post graduate classes of a university were asked to give their opinion towards the autonomous colleges. 190 of the Under-graduate and 210 of the post-graduate students favored the autonomous status.

Present the above fact in the form of frequency table and test, at 5% level, that opinions of Under-graduate and post-graduate student autonomous status of colleges are independent. (χ^2_{cal} = 3, do not reject H_0)

3. The number of married, unmarried and widow population in the three cities of Kathmandu Valley is obtained as below. Test whether the city and the marital status of the adult female population are associated.

	Married	Unmarried	Widow	Total
X	20	15	15	50
Y	30	20	25	75
Z	50	40	10	100
Total	100	75	50	225

(χ^2_{cal} = 15.994, reject H_0)

4. Explain the importance of inferential statistics in management and business science. A market analyst is reviewing three types of landed property in Singapore (terraces, detached, and semi-detached) that are owned by the three different household income groups (low, middle, and high). She organized the collected data into the following table:

Types of property	Household income range		
	Low	Middle	High
Terraces	15	57	80
Detached	32	84	20
Semi detached	8	13	47

Using 5% level of significant, help the analyst determine whether there is evidence of a significant relationship between type of landed property and the household income. (χ^2_{cal} = 71.266, reject H_0)