Chi-Square: Test for independence

Abhishek Negi
Student ID: 100004670
Msc. Computer science: Big data and AI (Cohort 2025)
SRH University
abhishek.negi53@gmail.com

Abhishek Negi
Student ID: 100004670
Msc. Computer science: Big data and AI (Cohort 2025)
SRH University
abhishek.negi53@gmail.com

Abhishek Negi
Student ID: 100004670
Msc. Computer science: Big data and AI (Cohort 2025)
SRH University
abhishek.negi53@gmail.com

April 30, 2025

Contents

1 Introduction

The Chi-Square Test of Independence is a statistical test used to determine whether there is a significant association between two categorical variables.

It tests the null hypothesis that the variables are independent — that is, the presence or level of one variable does not affect the other.

1.1 Background

The Chi-square test is a widely used non-parametric statistical test that was developed by Karl Pearson in 1900. It is used to assess whether observed data significantly differ from what would be expected under a specific hypothesis. Because it does not assume a normal distribution, it's especially useful for analyzing categorical data — data that can be divided into distinct groups or categories.

2 Methodology

2.1 Purpose:

To test whether two categorical variables are independent (i.e., not related) or associated in some way.

2.2 Basic Idea:

It compares the observed frequencies (actual data) in a contingency table with the expected frequencies (what we would expect if the variables were independent). If there's a large enough difference, the variables are likely associated.

2.3 Key Terminology:

- Observed frequency (O): The actual count in each cell of the contingency table.
- Expected frequency (E): The count you would expect if the variables were truly independent.
- Degrees of freedom (df) : $(rows - 1) * (columns - 1)$
- Chi-square statistic (χ^2) : $\sum \frac{(f_O - f_E)^2}{f_E}$

  f_O: observed frequencies
  f_E: expected frequencies

2.4 Steps to perform Chi-Square test for independence:

2.4.1 State the hypothesis:

- Null Hypothesis (H_0): The two variables are independent (no association).
- Alternative Hypothesis (H_A): The two variables are dependent (there is an association).

2.4.2 Prepare contingency table:

Organize your data into a contingency table. The rows and columns represent the two categorical variables you're testing for independence. Each cell in the table will contain the observed frequencies.
Example:

	Category A	Category B	Category C
Group 1	O_1	O_2	O_3
Group 2	O_4	O_5	O_6

Table 1: Contingency Table for Gender and Product Preference

2.4.3 Calculate expected frequency:

For each cell in the table, calculate the expected frequency assuming the null hypothesis is true. The formula for expected frequency E.
E for each cell is:

$$E = \frac{\text{(rows total)} * \text{(column total)}}{\text{grand total}} \tag{1}$$

Do this for every cell in your table.

2.4.4 Compute the Chi-Square statistics:

Use the following formula to compute the Chi-square statistic:

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

Where:

- O = Observed frequency
- E = Expected frequency

Sum this calculation for all the cells in your contingency table.

2.4.5 Define degree of freedom (df):

The degrees of freedom (df) for a Chi-square test for independence is given by:

$$df = (r - 1) * (c - 1) \tag{3}$$

Where:

- r = number of rows
- c = number of columns

2.4.6 Find the p-value or critical value:

- Using the Chi-square distribution table, find the critical value of Chi-square at the desired significance level (α, often 0.05) and degrees of freedom. The critical value corresponds to the threshold for rejecting the null hypothesis.
- Use a Chi-Square distribution table or software (like Excel, SPSS, or Python) to find the p-value corresponding to your χ^2 value and degrees of freedom.

2.4.7 Compare the statistic test with p-value or critical value:

- If p-value $\leq \alpha$ (typically 0.05), reject $H_O \rightarrow$ The variables are dependent.
- If p-value α, fail to reject $H_A \rightarrow$ The variables are independent.

  Or
- If the calculated Chi-square statistic is greater than the critical value from the Chi-square table, reject the null hypothesis (H_O).
- If the calculated Chi-square statistic is less than the critical value, fail to reject the null hypothesis.

2.4.8 Conclusion:

Based on the comparison in the previous step:

- If you reject H_O, conclude that there is a significant association between the two variables.
- If you fail to reject H_A , conclude that there is no significant association between the two variables.

3 Example

In this example we will use hypothetical data to calculate and better understand the chi-square test for independence.

Data:

Given below is an imaginary dataset which shows the product preferences (like or dislike) categorized based on gender (male or female):

	Like	Dislike	Total
Male	30	20	50
Female	40	10	50
Total	70	30	100

Table 2: Contingency Table for Gender and Product Preference

Step 1: Hypotheses

- Null Hypothesis (H_0): Gender and product preference are independent.
- Alternative Hypothesis (H_1): Gender and product preference are not independent.

Step 2: Calculate Expected Frequencies

The expected frequency for each cell is calculated using the Eq.(1):

For example, expected value for Male–Like:

$$E = \frac{50 \times 70}{100} = 35$$

Expected values for each cell:

- Male–Like: 35
- Male–Dislike: 15
- Female–Like: 35
- Female–Dislike: 15

Step 3: Compute the Chi-Square Statistic

Using Eq.(2):

$$\chi^2 = \frac{(30-35)^2}{35} + \frac{(20-15)^2}{15} + \frac{(40-35)^2}{35} + \frac{(10-15)^2}{15}$$

$$\chi^2 = \frac{25}{35} + \frac{25}{15} + \frac{25}{35} + \frac{25}{15}$$

$$\chi^2 \approx 0.714 + 1.667 + 0.714 + 1.667 = 4.762$$

Step 4: Degrees of Freedom

Using Eq.(3):

$$df = (r-1)(c-1) = (2-1)(2-1) = 1$$

Step 5: Determine Critical Value

At $\alpha = 0.05$ and $df = 1$, the critical value from the Chi-square table is **3.841**.

Step 6: Compare and Conclude

$$\chi^2 = 4.762 > 3.841 \Rightarrow \text{Reject } H_0$$

Step 7: Conclusion

There is a statistically significant association between gender and product preference.

4 Limitations of the Chi-Square Test for Independence

- **Requires large sample size:** The test may give misleading results when sample sizes are too small, especially if expected frequencies in any cell are less than 5.
- **Only for categorical data:** It cannot be used for continuous variables without converting them into categories, which can lead to loss of information.
- **Sensitive to sample distribution:** Uneven distribution among categories can distort the Chi-square value.
- **Does not indicate strength or direction:** The test only shows whether a relationship exists, not how strong or meaningful it is.
- **Assumes independence of observations:** The data must consist of independent observations; repeated or related measures violate this assumption.
- **Affected by grouping:** The way data is grouped or categorized can influence the results of the test.
- **Cannot handle sparse tables:** If too many cells have zero or very low expected frequencies, the validity of the test is compromised.

5 Result Interpretation

After performing the Chi-square test, the calculated Chi-square statistic is compared to the critical value from the Chi-square distribution table at the chosen significance level (usually $\alpha = 0.05$).

- **If** $\chi^2_{\text{calculated}} > \chi^2_{\text{critical}}$:
  - Reject the null hypothesis (H_0).
- **If** $\chi^2_{\text{calculated}} \leq \chi^2_{\text{critical}}$:
  - Fail to reject the null hypothesis (H_0).

6 General Conclusion Format

- **If the null hypothesis is rejected:**
  There is a statistically significant association between the two categorical variables. This suggests that the variables are **not independent**.
- **If the null hypothesis is not rejected:**
  There is no statistically significant association between the two categorical variables. This suggests that the variables are **independent**.

References

[1] Author Name, *Book Title*, Publisher, Year.