Chi-Square Test for Categorical Variables

Introduction

The chi-square test is a statistical method used to determine if there is a significant association between two categorical variables. This test is widely used in various fields, including social sciences, marketing, and healthcare, to analyze survey data, experimental results, and observational studies.

Concept

The chi-square test is a non-parametric statistical method that evaluates whether the observed frequencies in each category differ significantly from the expected frequencies—assuming no association between the variables.

The test is based on the chi-square distribution, which is a family of distributions defined by degrees of freedom (df). These distributions are right-skewed and vary depending on df. A chi-square distribution table lists critical values for given df and significance levels (α), which we use to assess if our computed test statistic is extreme enough to reject the null hypothesis.

Null Hypothesis and Alternative Hypothesis

The chi-square test involves formulating two hypotheses:

Null Hypothesis (H_0)(H0) - Assumes that there is no association between the categorical variables, implying that any observed differences are due to random chance.

Alternative Hypothesis (H_1)(H1) - Assumes that there is a significant association between the variables, indicating that the observed differences are not due to chance alone.

Formula

The **chi-square statistic** is calculated using the formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$χ2 = ∑ Ei(Oi−Ei)2

where

O_iOi is the observed frequency for category ii.

E_iEi is the expected frequency for category ii, calculated as:

$$E_i = \frac{(row\ total \times column\ total)}{grand\ total}$$Ei = grand total(row total×column total)

The sum is taken over all cells in the contingency table.

The calculated chi-square statistic is then compared to a critical value from the chi-square distribution table. This table provides critical values for different degrees of freedom (df)(df) and significance levels (α)(α).

The **degrees of freedom** for the test are calculated as:

$$df = (r - 1) \times (c - 1)$$df = (r − 1) × (c − 1)

where rr is the number of rows and cc is the number of columns in the table.

Chi-Square Distribution Table

A chi-square distribution table provides critical values that vary by degrees of freedom and the significance level (α). These values indicate the threshold beyond which the test statistic would be considered statistically significant.

For example:

df = 1, α = 0.05, the critical value is 3.841

If your calculated $\chi^2 > 3.841$, you reject H_0

If $\chi^2 \leq 3.841$, you fail to reject H_0

The higher the χ^2 value, the stronger the evidence against H_0.

Python Implementation Example

Below is a Python implementation using `scipy.stats and pandas`:

```
import pandas as pd
from scipy.stats import chi2_contingency
# Create the contingency table
data = [[20, 30],  # Male: [Like, Dislike]
        [25, 25]]  # Female: [Like, Dislike]
# Create a DataFrame for clarity
```

```
df = pd.DataFrame(data, columns=["Like", "Dislike"], index=["Male", "Female"])
# Perform the Chi-Square Test
chi2, p, dof, expected = chi2_contingency(df)
# Display results
print("Chi-square Statistic:", chi2)
print("Degrees of Freedom:", dof)
print("P-value:", p)
print("Expected Frequencies:\n", expected)
```

Output:

```
Chi-square Statistic: 1.008
Degrees of Freedom: 1
P-value: 0.3156
Expected Frequencies:[[22.5 27.5] [22.5 27.5]]
```

Interpretation: Since the p-value (0.3156) > 0.05, we fail to reject the null hypothesis—indicating no significant association.

Applications

1. **Market Research:** Analyzing the association between customer demographics and product preferences.
2. **Healthcare:** Studying the relationship between patient characteristics and disease incidence.
3. **Social Sciences:** Investigating the link between social factors (e.g., education level) and behavioral outcomes (e.g., voting patterns).
4. **Education:** Examining the connection between teaching methods and student performance.
5. **Quality Control:** Assessing the association between manufacturing conditions and product defects.

Practical Example - Weak Correlation

Suppose a researcher wants to determine if there is an association between gender (male, female) and preference for a new product (like, dislike). The researcher surveys 100 people and records the following data:

Category	Like	Dislike	Total
Male	20	30	50
Female	25	25	50
Total	45	55	100

Step 1: Calculate Expected Frequencies

Using the formula for expected frequencies:

$$E_{Male,Like} = \frac{(50 \times 45)}{100} = 22.5 \text{EMale,Like} = 100(50 \times 45) = 22.5$$

$$E_{Male,Dislike} = \frac{(50 \times 55)}{100} = 27.5 \text{EMale,Dislike} = 100(50 \times 55) = 27.5$$

$$E_{Female,Like} = \frac{(50 \times 45)}{100} = 22.5 \text{EFemale,Like} = 100(50 \times 45) = 22.5$$

$$E_{Female,Dislike} = \frac{(50 \times 55)}{100} = 27.5 \text{EFemale,Dislike} = 100(50 \times 55) = 27.5$$

Step 2: Compute Chi-Square Statistic

$$\chi^2 = \frac{(20-22.5)^2}{22.5} + \frac{(30-27.5)^2}{27.5} + \frac{(25-22.5)^2}{22.5} + \frac{(25-27.5)^2}{27.5} \chi 2 = 22.5(20-22.5)2 + 27.5(30-27.5)2 + 22.5(25-22.5)2 + 27.5(25-27.5)2$$

$$\chi^2 = \frac{(2.5)^2}{22.5} + \frac{(2.5)^2}{27.5} + \frac{(2.5)^2}{22.5} + \frac{(2.5)^2}{27.5} \chi 2 = 22.5(2.5)2 + 27.5(2.5)2 + 22.5(2.5)2 + 27.5(2.5)2$$

$$\chi^2 = \frac{6.25}{22.5} + \frac{6.25}{27.5} + \frac{6.25}{22.5} + \frac{6.25}{27.5} \chi 2 = 22.56.25 + 27.56.25 + 22.56.25 + 27.56.25$$

$$\chi^2 = 0.277 + 0.227 + 0.277 + 0.227 \chi 2 = 0.277 + 0.227 + 0.277 + 0.227$$

$$\chi^2 = 1.008 \chi 2 = 1.008$$

Step 3: Determine Degrees of Freedom

$$df = (2-1) \times (2-1) = 1\text{df} = (2-1) \times (2-1) = 1$$

Step 4: Interpret the Result

Using a chi-square distribution table, we compare the calculated chi-square value (1.008) with the critical value at one degree of freedom and a significance level (e.g., 0.05). The critical value, as determined from chi-square distribution tables, is approximately 3.841.

Since 1.008 < 3.841, we fail to reject the null hypothesis. Thus, there is no significant association between gender and product preference in this sample.

Practical Example - Strong Association

Consider a study investigating the relationship between smoking status (smoker, non-smoker) and the incidence of lung disease (disease, no disease). The researcher collects data from 200 individuals and records the following information:

Category	Disease	No Disease	Total
Smoker	50	30	80
Non-Smoker	20	100	120
Total	70	130	200

Step 1: Calculate Expected Frequencies

Using the formula for expected frequencies:

$$E_{Smoker,\,Disease} = \frac{(80 \times 70)}{200} = 28\text{ESmoker,Disease} = 200(80\times70) = 28$$

$$E_{Smoker,\,No\ Disease} = \frac{(80 \times 130)}{200} = 52\text{ESmoker,No Disease} = 200(80\times130) = 52$$

$$E_{Non-Smoker,\,Disease} = \frac{(120 \times 70)}{200} = 42\text{ENon−Smoker,Disease} = 200(120\times70) = 42$$

$$E_{Non-Smoker,\,No\ Disease} = \frac{(120 \times 130)}{200} = 78\text{ENon−Smoker,No Disease} = 200(120\times130) = 78$$

Step 2: Compute Chi-Square Statistic

$$\chi^2 = \frac{(50-28)^2}{28} + \frac{(30-52)^2}{52} + \frac{(20-42)^2}{42} + \frac{(100-78)^2}{78}\chi2 = 28(50−28)2 + 52(30−52)2 + 42(20−42)2 + 78(100−78)2$$

$$\chi^2 = \frac{(22)^2}{28} + \frac{(22)^2}{52} + \frac{(22)^2}{42} + \frac{(22)^2}{78}\chi2 = 28(22)2 + 52(22)2 + 42(22)2 + 78(22)2$$

$$\chi^2 = \frac{484}{28} + \frac{484}{52} + \frac{484}{42} + \frac{484}{78}\chi2 = 28484 + 52484 + 42484 + 78484$$

$$\chi^2 = 17.29 + 9.31 + 11.52 + 6.21\chi2 = 17.29 + 9.31 + 11.52 + 6.21$$

$$\chi^2 = 44.33\chi2 = 44.33$$

Step 3: Determine Degrees of Freedom

$$df = (2-1) \times (2-1) = 1\text{df} = (2-1) \times (2-1) = 1$$

Step 4: Interpret the Result

Using a chi-square distribution table, we compare the calculated chi-square value (44.33) with the critical value at one degree of freedom and a significance level (e.g., 0.05), approximately 3.841. Since 44.33 > 3.841, we reject the null hypothesis. This indicates a significant association between smoking status and the incidence of lung disease in this sample.

Conclusion

The chi-square test is a powerful tool for analyzing the relationship between categorical variables. By comparing observed and expected frequencies, researchers can determine if there is a statistically significant association, providing valuable insights in various fields of study.


Skills Network