

Question 1: Describe the procedure of Chi-Square Test of Independence.

Answer: The Chi-Square Test of Independence is a statistical method used to determine if there is a significant association between two categorical variables. This test is based on the comparison of observed frequencies in each category to the expected frequencies under a condition that the variables were independent. Steps to perform this test is as follows:

1. Stating the Hypotheses:

We have to state the Null and Alternate Hypothesis.

Null Hypothesis (H0): There is no association between the two categorical variables (i.e. they are independent).

Alternative Hypothesis (H1): There is an association between the two categorical variables (i.e. they are not independent).

2. Constructing the Contingency Table:

We have to organize the data into a contingency table, where rows represent the categories of one variable and columns represent the categories of the other variable. Each cell in the table represents the frequency count of observations falling into that cell.

3. Calculating the Expected Frequencies:

For each cell in the contingency table, we have to calculate the expected frequency under the assumption that the variables are independent. The expected frequency for a cell is calculated using the formula:

$$E_{ij} = (R_i * C_j) / N$$

where:

E_{ij} is the expected frequency for cell

R_i is the total count for row i

C_j is the total count for column j

N is the grand total of all observations.

4. Computing the Chi-Square Statistic:

We calculate the Chi-Square statistic using the formula:

$$\chi^2 = \sum ((O_{ij} - E_{ij})^2 / E_{ij})$$

where:

O_{ij} is the observed frequency for cell (i, j)

E_{ij} is the expected frequency for cell (i, j)

The summation is over all cells in the contingency table.

5. Determining the Degrees of Freedom:

The degrees of freedom for the test are calculated as:

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

where:

r is the number of rows in the contingency table

c is the number of columns in the contingency table

6. Finding the Critical Value and/or P-Value:

We have to compare the calculated Chi-Square statistic to the critical value from the Chi-Square distribution table based on the determined degrees of freedom and the chosen significance level (commonly $\alpha = 0.05$). Alternatively, we have to compute the p-value associated with the Chi-Square statistic.

7. Making a decision:

If the Chi-Square statistic exceeds the critical value or if the p-value is less than the significance level, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Example

Suppose we want to test whether there is an association between gender (Male, Female) and preference for a product (Like, Dislike). The data is as follows:

	Like	Dislike	Row Total
Male	20	10	30
Female	30	20	50
Column Total	50	30	80

1. Expected Frequencies:

$$E_{11} = (30 * 50) / 80 = 18.75$$

$$E_{12} = (30 * 30) / 80 = 11.25$$

$$E_{21} = (50 * 50) / 80 = 31.25$$

$$E_{22} = (50 * 30) / 80 = 18.75$$

2. Chi-Square Statistic:

$$\chi^2 = ((20 - 18.75)^2 / 18.75) + ((10 - 11.25)^2 / 11.25) + ((30 - 31.25)^2 / 31.25) + ((20 - 18.75)^2 / 18.75)$$

$$\chi^2 = (1.5625 / 18.75) + (1.5625 / 11.25) + (1.5625 / 31.25) + (1.5625 / 18.75)$$

$$\chi^2 = 0.0833 + 0.1389 + 0.0500 + 0.0833 = 0.3555$$

3. Degrees of Freedom:

$$df = (2 - 1) * (2 - 1) = 1 * 1 = 1$$

4. Critical Value / P-Value:

For $\alpha = 0.05$ and 1 degree of freedom, the critical value from the Chi-Square distribution table is 3.841.

Comparing $\chi^2 = 0.3555$ with 3.841, or we find the p-value using statistical software.

5. Decision:

Since χ^2 is less than 3.841, we fail to reject the null hypothesis. There is no significant association between gender and product preference at the 0.05 significance level.

This procedure allows us to determine if there is a significant association between two categorical variables using the Chi-Square Test of Independence.

Question 2: Implement logistic regression in R to predict whether a student will pass the exam based on their study hours and whether they attended a preparatory course.

Answer: