Lab 3 Chi-Square Test for Independence

https://courses.lumenlearning.com/odessa-introstats1-1/chapter/contingency-tables/Chi-Square Test for Independence

The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference. The sample problem at the end of the lesson considers this example.

When to Use Chi-Square Test for Independence

The test procedure described in this lesson is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

This approach consists of five steps:

- **Assumptions**
- **State The Hypotheses**
- **Formulate An Analysis Plan**
- **Analyze Sample Data**
- **Interpret Results.**

State the Hypotheses

Suppose that Variable A has *r* levels, and Variable B has *c* levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

H_o: Variable A and Variable B are independent.

H_a: Variable A and Variable B are not independent.

The alternative hypothesis is that knowing the level of Variable A *can* help you predict the level of Variable B.

Note: Support for the alternative hypothesis suggests that the variables are related; but the relationship is not necessarily causal, in the sense that one variable "causes" the other.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:

  $$DF = (r - 1) * (c - 1)$$

  where r is the number of levels for one catagorical variable, and c is the number of levels for the other categorical variable.

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r * c expected frequencies, according to the following formula.

  $$E_{r,c} = (n_r * n_c) / n$$

  where $E_{r,c}$ is the expected frequency count for level *r* of Variable A and level *c* of Variable B, n_r is the total number of sample observations at level r of Variable A, n_c is the total number of sample observations at level *c* of Variable B, and n is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

  $$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

where $O_{r,c}$ is the observed frequency count at level *r* of Variable A and level *c* of Variable B, and $E_{r,c}$ is the expected frequency count at level *r* of Variable A and level *c* of Variable B.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

Problem

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

	Voting Preferences			Row total
	Rep	Dem	Ind	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

H_o: Gender and voting preferences are independent.

H_a: Gender and voting preferences are not independent.

- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square test for independence.
- **Analyze sample data**. Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = (n_r * n_c) / n$$
$$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$$
$$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$$
$$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$$
$$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$$
$$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$$
$$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$$

$$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$
$$X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40 + (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$$
$$X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$$
$$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, n_r is the number of observations from level *r* of gender, n_c is the number of observations from level *c* of voting preference, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count when gender is level *r* and voting preference is level *c*, and $O_{r,c}$ is the observed frequency count when gender is level *r* voting preference is level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

$P(X^2 > 16.2) = 0.0003$.

- **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

Lab Assignment 3 20 marks

Ongoing competition in kaggle : https://www.kaggle.com/c/titanic/overview

Titanic Dataset Analysis: **This dataset contains data regarding the passengers of the Titanic.**

Qns 1-6 -10 marks, Qn 7,8 -10 marks

1.Read the Titanic data set into a data frame called “titanic”.

2. Preprocess the data

3. Count the total number of passengers on board the Titanic.

4. Count the number of passengers who survived the sinking of the Titanic.

5. Measure the percentage of passengers who survived the sinking of the Titanic.

6. Count the number of passengers based on gender

7. Run a Pearson’s Chi-squared test to test the following hypothesis:

Hypothesis: The proportion of females onboard who survived the sinking of the Titanic was higher than the proportion of males onboard who survived the sinking of the Titanic.

8. Inference based on test.