Chi-Square Test for Association and Independence

by Sophia

WHAT'S COVERED

This tutorial will cover the chi-square test of independence. Our discussion breaks down as follows:

1. The Chi-Square Test for Association/Independence

The **chi-square test for association** is sometimes called a chi-square test of independence. This is a type of hypothesis test to test if there is no association between multiple categorical variables in a *single population*.

As with any chi-square test, you must follow these steps:

STEP BY STEP

Step 1: State the null and alternative hypotheses.
Step 2: Check the conditions.
Step 3: Calculate the test-statistic and p-value
Step 4: Compare your test statistic to your chosen critical value, or your p-value to your chosen significance level. Based on how they compare, state a decision about the null hypothesis and conclusion in the context of the problem.

Recall that the conditions for a chi-square test are:

- The data represent a simple random sample from the population.
- The observations should be sampled independently from the population, and the population is at least 10 times sample size condition, which is called the "10% of the population" condition.
- The expected counts have to be at least 5. We have to ensure that the sample size is large, which is similar to the conditions for checking normality in other hypothesis tests.

EXAMPLE Suppose 335 students of different backgrounds (rural, suburban, and urban schools) were asked to pick one thing about school that was most important to them: getting good grades, being popular, or being good at sports. Here is the distribution of responses:

	School Locations		
Goal	**Rural**	**Suburban**	**Urban**
Grades	57	87	24
Popular	50	42	6
Sports	42	22	5

The question is, does there appear to be an association between the geographic location of the school and the answer choice to the question (the goal)? This is an ideal time to run a chi-square test for association or independence. This can tell you if the distribution of goals (grades, popular, and sports) differ significantly for each school location. Are they associated or are they independent?

Step 1: State the null and alternative hypotheses.

In the null hypothesis, you're going to say that school location and goal are independent. That is, they do *not* have an association with each other. The alternative hypothesis is that they do have an association with each other. At least one of these distributions--grades, popularity, and sports--is different for suburban, urban, or rural verus the others. Also, you can choose a significance level of 0.05.

- H_0: The school locations and goals are independent.
- H_a: The school locations and goals are associated.
- α: 0.05

Step 2: Check the conditions.

For the test of independence, the conditions and the way that chi-square and p-value are calculated are the same as in a test of homogeneity. We first need to find the expected value for each cell to ensure that the condition is met.

Remember, the expected value is equal to that particular cell's row total, times its column total, divided by the grand total for all the cells.

FORMULA TO KNOW

Expected Value for Cell in Chi-Square Test for Association/Independence

$$Expected\ Value\ for\ Cell = \frac{(Row\ Total)(Column\ Total)}{Grand\ Total}$$

For example, if we wanted the expected value for "Grades" and "Rural", we would multiply the row total for

"Grades" with the column total for "Rural", and divide by the total values in the table.

$$\textit{Expected Value for "Grades" and "Rural"} = \frac{(168)(149)}{335} = 74.72$$

For the row with "Grades", there was a total of 57 plus 87 plus 24, or 168 students. For "Rural", there was a total of 149 students. We were told at the beginning there were a total of 335 students, however, we could also add up all the values in the table to get this same value.

We can continue using this formula for each cell and get the expected table of results:

Observed

	School Locations		
Goal	**Rural**	**Suburban**	**Urban**
Grades	57	87	24
Popular	50	42	6
Sports	42	22	5

Expected

	School Locations		
Goal	**Rural**	**Suburban**	**Urban**
Grades	74.72	75.73	17.55
Popular	43.59	44.17	10.24
Sports	30.69	31.10	7.21

What you are interested in is whether or not all the expected counts are at least 5. The smallest one is 7.21, so the conditions are met.

Step 3: Calculate the test-statistic and p-value.

Using technology, we can find that the chi-square statistic is equal to 18.564, which is big.

To find the corresponding p-value, we first need to find the degrees of freedom. The degrees of freedom can be found by multiplying the number of rows minus one times the value of the number of columns minus one.

FORMULA TO KNOW

Chi-Square Test Degrees of Freedom

$$Degrees\ of\ freedom = (row\ total - 1)(column\ total - 1)$$

In this case, there were three rows (grades, popular, and sports) and three columns (rural, suburban, and urban):

$$Degrees\ of\ freedom = (3-1)(3-1) = (2)(2) = 4$$

In this case, the degrees of freedom is going to be equal to four. Using technology and plugging in the chi-square statistic of 18.564 and 4 degrees of freedom, the p-value can be obtained and we get a very small p-value of 0.001.

Step 4: Compare your test statistic to your chosen critical value, or your p-value to your chosen significance level. Based on how they compare, state a decision about the null hypothesis and conclusion in the context of the problem.

You need to link your p-value to a decision about the null hypothesis. Since the p-value is smaller than 0.05, you reject the null hypothesis in favor of the alternative and conclude that there is an association between the two categorical variables of school location and goal.

TERM TO KNOW

Chi-Square Test for Association/Independence

A hypothesis test that tests whether two qualitative variables have an association or not.

SUMMARY

The chi-square test of independence tests whether two qualitative variables have an association or not, so it's sometimes called the chi-square test of association. The expected value for each cell is equal to that particular cell's row total, times its column total, divided by the grand total for all the cells.

Good luck!

Source: THIS TUTORIAL WAS AUTHORED BY JONATHAN OSTERS FOR SOPHIA LEARNING. PLEASE SEE OUR TERMS OF USE.

TERMS TO KNOW

Chi-Square Test of Independence/Association

A hypothesis test that tests whether two qualitative variables have an association or not.

FORMULAS TO KNOW

Chi-square Degrees of Freedom

$$degrees\ of\ freedom = (row\ total - 1)(column\ total - 1)$$