“What is the Chi-Square Test of Independence and how is it used in SAS Tutorials?

Authored by

stats writer

June 24, 2024

The Chi-Square Test of Independence is a statistical test used to determine whether there is a significant relationship between two categorical variables. It is commonly used to analyze data from contingency tables, where the rows represent one variable and the columns represent the other. This test calculates the expected frequencies for each category based on the assumption of independence, and then compares them to the observed frequencies in the data. The resulting p-value indicates the likelihood of obtaining the observed data if the variables were truly independent. In SAS Tutorials, the Chi-Square Test of Independence is used to assess the relationship between two categorical variables and determine whether this relationship is significant or due to chance. This allows users to make informed decisions and draw meaningful conclusions from their data.

Chi-Square Test of Independence

The Chi-Square Test of Independence determines whether there is an association between categorical variables (i.e., whether the variables are independent or related). It is a nonparametric test.

This test is also known as:

Chi-Square Test of Association.

This test utilizes a contingency table to analyze the data. A contingency table (also known as a *cross-tabulation*, *crosstab*, or *two-way table*) is an arrangement in which data is classified according to two categorical variables. The categories for one variable appear in the rows, and the categories for the other variable appear in columns. Each variable must have two or more categories. Each cell reflects the total count of cases for a specific pair of categories.

There are several tests that go by the name "chi-square test" in addition to the Chi-Square Test of Independence. Look for context clues in the data and research question to make sure what form of the chi-square test is being used.

Common Uses

The Chi-Square Test of Independence is commonly used to test the following:

Statistical independence or association between two categorical variables.

The Chi-Square Test of Independence can only compare categorical variables. It cannot make comparisons between continuous variables or between categorical and continuous variables. Additionally, the Chi-Square Test of Independence only assesses *associations* between categorical variables, and can not provide any inferences about causation.

If your categorical variables represent "pre-test" and "post-test" observations, then the chi-square test of independence **is not appropriate**. This is because the assumption of the independence of observations is violated. In this situation, McNemar's Test is appropriate.

Data Requirements

Your data must meet the following requirements:

Two categorical variables.Two or more categories (groups) for each variable.Independence of observations.
There is no relationship between the subjects in each group.The categorical variables are not "paired" in any way (e.g. pre-test/post-test observations).Relatively large sample size.
Expected frequencies for each cell are at least 1.Expected frequencies should be at least 5 for the majority (80%) of the cells.

Hypotheses

The null hypothesis (H0) and alternative hypothesis (H1) of the Chi-Square Test of Independence can be expressed in two different but equivalent ways:

H0: " is independent of "
H1: " is not independent of "

OR

*H*0: " is not associated with "
H1: " is associated with "

Data Set-Up

Your dataset should have the following structure:

Each case (row) represents a subject, and each subject appears once in the dataset, represented in columns. That is, each row represents an observation from a unique subject.The dataset contains at least two nominal categorical variables (string or numeric). The categorical variables used in the test must have two or more categories; they should also not have too many categories.

	ID Number	Class rank	Gender	Are you an athlete?
1	20183	.	Male	Non-athlete
2	20230	Freshman	Male	Athlete
3	20243	Junior	Female	Non-athlete
4	20248	Freshman	.	Non-athlete
5	20255	Sophomore	Female	Non-athlete
6	20278	.	Male	Non-athlete
7	20389	.	Male	Non-athlete
8	20402	Sophomore	Male	Non-athlete
9	20531	Freshman	Male	Athlete
10	20615	Freshman	Female	Non-athlete
11	20626	Sophomore	Female	Non-athlete
. . .				
425	49386	Sophomore	Female	Non-athlete
426	49445	Sophomore	Male	Athlete
427	49572	Senior	Female	Non-athlete
428	49688	.	Male	Non-athlete
429	49806	Junior	Female	Non-athlete
430	49821	Junior	Female	Athlete
431	49838	Freshman	Male	Athlete
432	49854	Junior	Male	Non-athlete
433	49879	Sophomore	Male	Athlete
434	49931	Junior	Male	Athlete
435	49947	Freshman	Female	Athlete

Test Statistic

The test statistic for the Chi-Square Test of Independence is denoted $X2$, and is computed as:

$$ chi^{2} = sum_{i=1}^{R}{sum_{j=1}^{C}{frac{(o_{ij} - e_{ij})^{2}}{e_{ij}}}} $$

where

(o_{ij}) is the observed cell count in the ith row and jth column of the table

(e_{ij}) is the expected cell count in the ith row and jth column of the table, computed as

$$ e_{ij} = frac{mathrm{ textrm{row } mathit{i}} textrm{ total} * mathrm{textrm{col } mathit{j}} textrm{ total}}{textrm{grand total}} $$

The quantity (oij - eij) is sometimes referred to as the *residual* of cell (i, j), denoted (r_{ij}).

The calculated $X2$ value is then compared to the critical value from the $X2$ distribution table with

degrees of freedom df = (R - 1)(C - 1) and chosen confidence level. If the calculated $X2$ value > critical $X2$ value, then we reject the null hypothesis.

Run a Chi-Square Test of Independence with PROC FREQ

The general form is

```
PROC FREQ data=dataset-name;
TABLE rowVar*colVar / CHISQ;
RUN;
```

The CHISQ option is added to the TABLES statement after the slash (/) character.

Many of PROC FREQ's most useful options have been covered in the tutorials on Frequency Tables and Crosstabs, but there are several additional options that can be useful when conducting a chi-square test of independence:

`EXPECTED`
Adds expected cell counts to the cells of the crosstab table.`DEVIATION`
Adds deviation values (i.e., observed minus expected values) to the cells of the crosstab table.

Example: Chi-square Test for 2x2 Table

Problem Statement

Let's continue the row and column percentage example from the Crosstabs tutorial, which described the relationship between the variables *RankUpperUnder* (upperclassman/underclassman) and *LivesOnCampus* (lives on campus/lives off-campus). Recall that the column percentages of the crosstab appeared to indicate that upperclassmen were less likely than underclassmen to live on campus:

The proportion of underclassmen who live off campus is 34.8%, or 79/227.The proportion of underclassmen who live on campus is 65.2%, or 148/227.The proportion of upperclassmen who live off campus is 94.4%, or 152/161.The proportion of upperclassmen who live on campus is 5.6%, or 9/161.

Suppose that we want to test the association between class rank and living on campus using a Chi-Square Test of Independence (using $\alpha = 0.05$).

Syntax

```
PROC FREQ DATA=work.sample;
TABLE RankUpperUnder*LiveOnCampus / CHISQ EXPECTED DEVIATION NOROW NOCOL NOPERCENT;
RUN;
```

Output

The first table in the output is the crosstabulation. If you included the `EXPECTED` and `DEVIATION` options in your syntax, you should see the following:

Frequency Expected Deviation	Table of RankUpperUnder by LiveOnCampus		
RankUpperUnder(Class Rank (binary))	LiveOnCampus(Do you live on campus?)		
	Off-campus	On-campus	Total
Underclassmen	79 135.15 -56.15	148 91.853 56.147	227
Upperclassmen	152 95.853 56.147	9 65.147 -56.15	161
Total	231	157	388
Frequency Missing = 47			

With the Expected Count values shown, we can confirm that all cells have an expected value greater than 5.

	Off-Campus	On-Campus	Total
Underclassman	Row 1, column 1 $$ o_{mathrm{11}} = 79 $$ $$ e_{mathrm{11}} = frac{227*231}{388} = 135.15 $$ $$ r_{mathrm{11}} = 79 - 135.147 = -56.15 $$	Row 1, column 2 $$ o_{mathrm{12}} = 148 $$ $$ e_{mathrm{12}} = frac{227*157}{388} = 91.853 $$ $$ r_{mathrm{12}} = 148 - 91.853 = 56.147 $$	row 1 total = 227

	Off-Campus	On-Campus	Total
Upperclassmen	Row 2, column 1 $$ o_{mathrm{21}} = 152 $$ $$ e_{mathrm{21}} = frac{161*231}{388} = 95.853 $$ $$ r_{mathrm{21}} = 152 - 95.853 = 56.147 $$	Row 2, column 2 $$ o_{mathrm{22}} = 9 $$ $$ e_{mathrm{22}} = frac{161*157}{388} = 65.147 $$ $$ r_{mathrm{22}} = 9 - 65.147 = -56.15 $$	row 2 total = 161
Total	col 1 total = 231	col 2 total = 157	grand total = 388

These numbers can be plugged into the chi-square test statistic formula:

$$ chi^{2} = sum_{i=1}^{R}{sum_{j=1}^{C}{frac{(o_{ij} - e_{ij})^{2}}{e_{ij}}}} = frac{(-56.15)^{2}}{135.15} + frac{(56.147)^{2}}{91.853} + frac{(56.147)^{2}}{95.853} + frac{(-56.15)^{2}}{65.147} = 138.926 $$

We can confirm this computation with the results in the table labeled **Statistics for Table of RankUpperUnder by LiveOnCampus**:

Statistics for Table of RankUpperUnder by LiveOnCampus

Statistic	DF	Value	Prob
Chi-Square	1	138.9260	<.0001
Likelihood Ratio Chi-Square	1	160.8998	<.0001
Continuity Adj. Chi-Square	1	136.4627	<.0001
Mantel-Haenszel Chi-Square	1	138.5679	<.0001
Phi Coefficient		-0.5984	
Contingency Coefficient		0.5135	
Cramer's V		-0.5984	

The row of interest here is **Chi-Square**.

The value of the test statistic is 138.926.Because the crosstabulation is a 2x2 table, the degrees of freedom (df) for the test statistic is $$ df = (R - 1)*(C - 1) = (2 - 1)*(2 - 1) = 1 $$.The corresponding p-value of the test statistic is so small that it is presented as $p < 0.001$.

Decision and Conclusions

Since the p-value is less than our chosen significance level $\alpha = 0.05$, we can reject the null hypothesis, and conclude that there is an association between class rank and whether or not students live on-campus.

Based on the results, we can state the following:

There was a significant association between class rank and living on campus ($X2(1) = 138.9$, $p < .001$).

Tutorial Feedback