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 or more 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:

1. Two categorical variables.
2. Two or more categories (groups) for each variable.
3. 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).
4. 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 (H_0) and alternative hypothesis (H_1) of the Chi-Square Test of Independence can be expressed in two different but equivalent ways:

H_0: "[*Variable 1*] is independent of [*Variable 2*]"
H_1: "[*Variable 1*] is not independent of [*Variable 2*]"

OR

H_0: "[*Variable 1*] is not associated with [*Variable 2*]"
H_1*:* "[*Variable 1*] is associated with [*Variable 2*]"

Test Statistic

The test statistic for the Chi-Square Test of Independence is denoted X^2, and is computed as:

$$\chi^2=\sum^{i=1R}\sum^{j=1C}(o^{ij}-e^{ij})^2e^{ij}\chi 2=\sum i=1R\sum j=1C(oij-eij)2eij$$

where

o^{ij}oij is the observed cell count in the i^{th} row and j^{th} column of the table

e^{ij}eij is the expected cell count in the i^{th} row and j^{th} column of the table, computed as

e^{ij}=row i total∗col j totalgrand totaleij=row i total∗col j totalgrand total

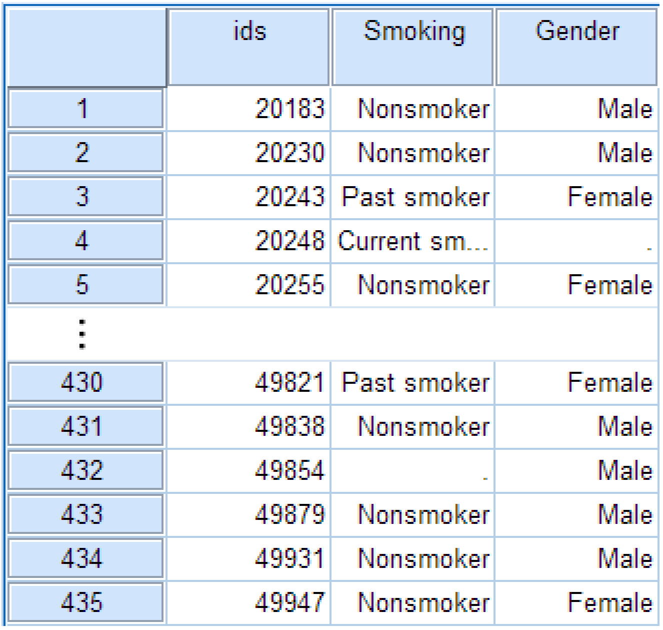
The quantity (o^{ij} - e^{ij}) is sometimes referred to as the *residual* of cell (i, j), denoted r^{ij}rij.

The calculated X^2 value is then compared to the critical value from the X^2 distribution table with degrees of freedom $df = (R - 1)(C - 1)$ and chosen confidence level. If the calculated X^2 value > critical X^2 value, then we reject the null hypothesis.

Data Set-Up

There are two different ways in which your data may be set up initially. The format of the data will determine how to proceed with running the Chi-Square Test of Independence. At minimum, your data should include two categorical variables (represented in columns) that will be used in the analysis. The categorical variables must include at least two groups. Your data may be formatted in either of the following ways:

IF YOU HAVE THE RAW DATA (EACH ROW IS A SUBJECT):

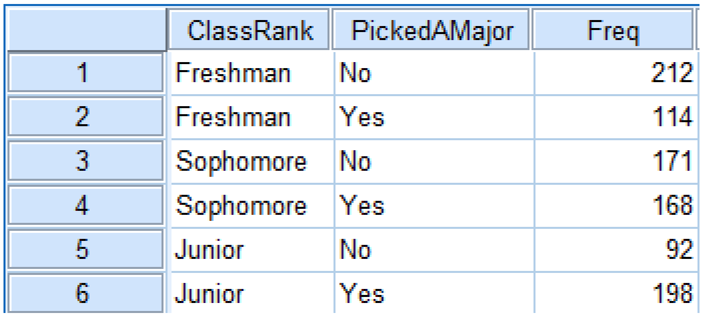


	ids	Smoking	Gender
1	20183	Nonsmoker	Male
2	20230	Nonsmoker	Male
3	20243	Past smoker	Female
4	20248	Current sm...	.
5	20255	Nonsmoker	Female
⋮			
430	49821	Past smoker	Female
431	49838	Nonsmoker	Male
432	49854	.	Male
433	49879	Nonsmoker	Male
434	49931	Nonsmoker	Male
435	49947	Nonsmoker	Female

- Cases represent subjects, and each subject appears once in the dataset. 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.

IF YOU HAVE FREQUENCIES (EACH ROW IS A COMBINATION OF FACTORS):

An example of using the chi-square test for this type of data can be found in the Weighting Cases tutorial.



	ClassRank	PickedAMajor	Freq
1	Freshman	No	212
2	Freshman	Yes	114
3	Sophomore	No	171
4	Sophomore	Yes	168
5	Junior	No	92
6	Junior	Yes	198

- Cases represent the combinations of categories for the variables.
  - Each row in the dataset represents a distinct combination of the categories.
  - The value in the "frequency" column for a given row is the number of unique subjects with that combination of categories.
- You should have three variables: one representing each category, and a third representing the number of occurrences of that particular combination of factors.
- Before running the test, you must activate Weight Cases, and set the frequency variable as the weight.