Perform a Chi-Square Test of Independence in SAS

Authored by

Mohammed looti

November 1, 2025

The Chi-Square Test of Independence is a cornerstone statistical procedure utilized to rigorously assess whether a statistically significant association exists between two categorical variables within a defined population. This non-parametric test is essential across diverse fields, including the social sciences, market analysis, and epidemiology, where researchers frequently analyze how frequencies are distributed across different groups.

At its core, the test evaluates the principle of independence: if the variables are truly independent, the distribution of one variable's categories should remain uniform regardless of the category of the second variable. Conversely, a dependency implies that the distribution will shift significantly across categories, suggesting a relationship. This comprehensive guide details the complete workflow for conducting this vital test using the powerful statistical analysis system, SAS.

Understanding the Theoretical Basis of the Chi-Square Test

Before implementing the analysis in SAS, it is crucial to grasp the fundamental principles driving the Chi-Square Test. The test operates by comparing the observed frequencies--the actual data collected from our sample--with the expected frequencies--the counts we would anticipate if the two variables were perfectly independent. The magnitude of the difference between these two sets of frequencies determines the resulting Chi-Square test statistic.

The calculated test statistic follows a Chi-Square distribution, which enables the calculation of an associated p-value. The p-value represents the probability of observing data as extreme as, or more extreme than, the current data, assuming that the null hypothesis (H0) of independence is correct. If this p-value is sufficiently small, it suggests that the observed differences are highly unlikely to be the result of random chance alone, leading us to reject the assumption of independence.

This statistical procedure is explicitly designed for data summarized and presented in a contingency table (or cross-tabulation), where observations are classified based on two distinct factors. The overarching objective is to statistically evaluate whether the categorization of one variable systematically influences the probability of being categorized a certain way in the other variable.

Defining the Research Scenario and Data Structure

For our practical demonstration, we will address a common social science inquiry: Is there an association between a voter's gender and their political party preference? To investigate this question, we hypothesize that a survey was administered to a random sample consisting of 500 registered voters.

The two categorical variables under scrutiny are **Gender** (categorized as Male or Female) and

Political Party Preference (categorized as Republican, Democrat, or Independent). The survey results are summarized in the contingency table below, which displays the raw frequency counts for every unique combination of categories. This highly organized format is perfectly suited for direct input and subsequent analysis within the SAS environment.

The table below represents the observed frequencies collected from the 500 survey respondents:

	Republican	Democrat	Independent	Total
Male	120	90	40	250
Female	110	95	45	250
Total	230	185	85	500

The subsequent steps demonstrate the necessary SAS programming required to transform this summarized frequency data into a format suitable for formal hypothesis testing, ultimately allowing us to determine if the observed distribution patterns are statistically significant.

Step 1: Preparing and Loading the Data in SAS

The fundamental requirement for any analysis in SAS is correctly structuring the data into a SAS dataset. Since our data is already aggregated in a contingency table, we must use the **DATA step** in conjunction with the **INPUT** and **DATALINES** statements to input the data directly. We need to define three distinct variables: two character variables to hold the categories (Gender and Party) and one numeric variable (Count) to hold the frequency, or cell count, of observations for that combination.

The inclusion of the **$** symbol immediately following the variable name in the **INPUT** statement is vital; it explicitly designates Gender and Party as character variables, which is necessary because they are non-numeric labels. The subsequent data lines then systematically input the cell counts from the table above into our new dataset, which we name my_data.

The following code block executes the data creation process and uses the PROC PRINT procedure to display the resulting data table structure in the SAS output window for verification:

```
/*create dataset*/
data my_data;
input Gender $ Party $ Count;
datalines;
Male Rep 120
Male Dem 90
Male Ind 40
```

```
Female Rep 110
Female Dem 95
Female Ind 45
;
run;

/*print dataset*/
proc print data=my_data;
```

Running the script confirms that the data has been loaded successfully. The output shows six rows, representing the six unique combinations of Gender and Party Preference, each associated with the correct frequency count.

Obs	Gender	Party	Count
1	Male	Rep	120
2	Male	Dem	90
3	Male	Ind	40
4	Female	Rep	110
5	Female	Dem	95
6	Female	Ind	45

Step 2: Executing the Chi-Square Analysis using PROC FREQ

The procedure required to perform the Chi-Square Test of Independence in SAS is PROC FREQ. This procedure is specifically optimized for calculating frequencies, generating cross-classification tables, and computing various tests of association, including the necessary Chi-Square statistic.

Within the **PROC FREQ** block, the **TABLES** statement is employed to define the variables for cross-tabulation. We specify the relationship between Gender and Party using the standard SAS syntax: Gender*Party. Following the slash (/), the **CHISQ** option is the critical directive that instructs SAS to calculate the Pearson Chi-Square statistic, the likelihood ratio Chi-Square, and related measures of association.

A crucial adjustment is necessary because we are inputting summarized data (frequencies) rather than a list of 500 individual observations. We must include the **WEIGHT** statement. The command WEIGHT Count; explicitly informs SAS that the variable Count holds the frequency of occurrence for each row. This ensures that the statistical calculations correctly use the total sample size (N=500) rather than incorrectly treating the six input rows as the entire dataset.

```
/*perform Chi-Square Test of Independence*/
proc freq data=my_data;
tables Gender*Party / chisq;
weight Count;
run;
```

Interpreting the SAS Output and Statistical Results

Execution of the PROC FREQ code generates a detailed output, including the full contingency table and a section dedicated to tests of association. For our hypothesis test, we focus exclusively on the table labeled "Statistics for Table of Gender by Party," which contains the summary of the Chi-Square results.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Gender by Party

Gender	Party			
	Dem	Ind	Rep	Total
Female	95 19.00 38.00 51.35	45 9.00 18.00 52.94	110 22.00 44.00 47.83	250 50.00
Male	90 18.00 36.00 48.65	40 8.00 16.00 47.06	120 24.00 48.00 52.17	250 50.00
Total	185 37.00	85 17.00	230 46.00	500 100.00

Statistics for Table of Gender by Party

Statistic	DF	Value	Prob
Chi-Square	2	0.8640	0.6492
Likelihood Ratio Chi-Square	2	0.8644	0.6491
Mantel-Haenszel Chi-Square	1	0.5464	0.4598
Phi Coefficient		0.0416	
Contingency Coefficient		0.0415	
Cramer's V		0.0416	

Sample Size = 500

The essential components for making a statistical decision regarding the independence of the

variables are the Chi-Square statistic and its corresponding p-value, which are found in the Pearson Chi-Square row:

Chi-Square Test Statistic: **0.8640**
Degrees of Freedom (DF): **2**
Corresponding P-value: **0.6492**

To formally conclude the test, we must first articulate the statistical hypotheses being tested:

H0 (Null Hypothesis): Gender and Political Party Preference are independent. There is no statistically significant association between them in the voter population.
HA (Alternative Hypothesis): Gender and Political Party Preference are *not* independent. There is a statistically significant association between the two variables.

Our decision rule relies on comparing the calculated p-value to a predetermined level of statistical significance, commonly denoted as alpha (α), which we set at 0.05. If the p-value is less than 0.05, we reject the null hypothesis in favor of the alternative.

In this specific analysis, the resulting p-value (0.6492) is substantially larger than our established significance level of 0.05. Consequently, we must fail to reject the null hypothesis (H0).

Conclusion and Practical Implications

The statistical decision to fail to reject the null hypothesis leads to a clear and definitive conclusion: Based on the evidence gathered from this sample, there is insufficient statistical evidence to assert a significant association between gender and political party preference. The minor numerical differences observed in the contingency table are likely attributed to expected random sampling variability rather than reflecting a genuine underlying relationship within the population of registered voters.

In practical terms, this result suggests that knowing a voter's gender does not provide a statistically reliable basis for predicting their political party preference. For the population represented by our sample, these two factors appear to behave as statistically independent characteristics. This reinforces the idea that caution is required when interpreting raw count differences without formal hypothesis testing.

Additional Resources for Deeper Chi-Square Understanding

To further enhance proficiency in applying and interpreting the Chi-Square Test of Independence, especially in complex datasets, consider exploring these related statistical topics that provide essential context regarding assumptions and limitations:

Calculating Expected Frequencies and verifying the minimum cell count assumptions.
Understanding the Degrees of Freedom (DF) Calculation in a cross-classification table.
Using the Likelihood Ratio Chi-Square Test as an alternative, particularly when dealing with small observed frequencies.