

How to Easily Perform a Chi-Square Test of Independence in SAS

Authored by
stats writer

December 1, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Easily Perform a Chi-Square Test of Independence in SAS*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=103456>

The Chi-Square Test of Independence is a foundational statistical method used to determine if a significant relationship exists between two categorical variables. When conducting this analysis within SAS, the process involves leveraging built-in procedures to first generate the necessary cross-tabulation table and then calculate the critical test statistic. This approach allows researchers to rigorously test their hypotheses about variable associations using industry-standard statistical software.

To execute the analysis in SAS, we primarily utilize the PROC FREQ procedure. This powerful tool generates the contingency table, which is the necessary foundation for the Chi-Square calculation. Furthermore, by specifying the appropriate options within PROC FREQ, we can instruct SAS to automatically compute the Chi-Square test statistic and, most importantly, the corresponding p-value. The magnitude of the p-value is the ultimate determinant in assessing the statistical significance of the observed association between the variables under study.

Understanding the output of this procedure is essential for drawing accurate conclusions. If the calculated p-value falls below a predetermined significance level (commonly 0.05), we reject the null hypothesis, indicating a statistically significant association between the two variables. Conversely, if the p-value is high, we conclude that there is insufficient evidence to suggest dependence. This detailed guide demonstrates a step-by-step example for performing a Chi-Square Test of Independence efficiently in SAS.

The Chi-Square Test of Independence is formally applied to assess whether the distribution of one variable is independent of the distribution of a second variable.

The following detailed scenario illustrates the complete process, from data creation to interpretation, required to execute the Chi-Square Test of Independence in SAS.

Example: Testing Association Between Gender and Political Preference

Consider a research question focused on understanding potential links between demographic characteristics and political behavior. Specifically, we aim to determine whether or not a voter's **gender** is statistically associated with their **political party preference**. This constitutes a classic application for the Chi-Square Test of Independence, as both variables are categorical.

To gather the necessary evidence, we conducted a systematic random sample of 500 eligible voters. Each participant was surveyed regarding their gender (Male or Female) and their primary political affiliation (Republican, Democrat, or Independent). The results of this comprehensive survey were compiled into a contingency table format, summarizing the cross-classification of responses.

The table below presents the observed frequencies gathered from the 500 respondents:

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

Setting Up the Analysis in SAS

To proceed with the statistical testing, the raw frequency data must first be correctly structured and loaded into a SAS dataset. This critical initial phase ensures that the statistical procedures can accurately read and process the observed counts for each combination of gender and political party. We will follow a two-step approach: first, data creation and verification, and second, execution of the frequency procedure for the test.

The following sequence of steps details the necessary SAS programming required to perform the Chi-Square Test of Independence and evaluate if gender is associated with political party preference based on the survey data.

Step 1: Creating and Verifying the SAS Dataset

The initial requirement is to create a well-structured dataset in SAS that accurately represents the frequency table above. Since the data is already aggregated (we have counts rather than one row per voter), we need to use the DATALINES statement to input the variables `Gender`, `Party`, and `Count`, where `Count` holds the frequency for that specific category combination.

The `INPUT` statement defines the structure of the variables, noting that `Gender` and `Party` are character variables (denoted by the \$). The `PROC PRINT` command is then used immediately after the data step to verify the successful creation and accurate entry of the dataset, ensuring the data is ready for analysis.

```
/*create dataset: my_data holds the aggregated frequency counts*/
```

```
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 for verification*/
proc print data=my_data;
run;
```

Upon executing the data step and the print procedure, the resulting output confirms the dataset structure and data accuracy, aligning perfectly with the original contingency table:

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: Performing the Chi-Square Test of Independence using PROC FREQ

With the data successfully loaded, the next step involves running the primary analysis using `PROC FREQ`. This procedure is uniquely suited for analyzing categorical data and generating contingency tables. To ensure the procedure uses the aggregated counts rather than assuming one observation per row, the `WEIGHT` statement is crucial; it tells SAS that the variable `Count` represents the frequency of occurrence for each combination.

The `TABLES` statement specifies the cross-tabulation desired--`Gender*Party`--and critically, the `CHISQ` option is added to request the output of the Chi-Square statistics necessary for the Test of Independence. This single procedure handles both the table generation and the statistical test computation simultaneously.

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

Executing this code block generates extensive output, including the detailed frequency table,

expected counts, and, most importantly, the summary statistics for the Chi-Square tests.

The FREQ Procedure

Table of Gender by Party				
Gender	Party			Total
	Dem	Ind	Rep	
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

Interpreting the SAS Output

The output generated by `PROC FREQ` provides several statistics, but for the Chi-Square Test of Independence, we focus primarily on two key values found under the "Statistics for Table of Gender by Party" section:

Chi-Square Test Statistic (Pearson): **0.8640**. This value quantifies the difference between the observed cell frequencies and the frequencies that would be expected if the variables were truly independent.

Corresponding Probability (p-value): **0.6492**. This is the probability of observing a Chi-Square statistic as extreme as 0.8640 (or more extreme) if the null hypothesis were true.

The degrees of freedom (df) associated with this test are calculated as (Rows - 1) * (Columns - 1),

which in this case is $(2 - 1) * (3 - 1) = 2$.

Formulating the Hypotheses and Drawing Conclusions

Before making a final decision, it is necessary to formally state the statistical hypotheses being tested. The Chi-Square Test of Independence operates under the following framework:

H0 (Null Hypothesis): The two variables (Gender and Political Party Preference) are statistically **independent**. There is no association between them in the population.

HA (Alternative Hypothesis): The two variables are **not** independent. There is a statistically significant association between Gender and Political Party Preference.

To assess these hypotheses, we compare the calculated p-value (0.6492) against the standard significance level (alpha = 0.05). The decision rule dictates that we only reject the null hypothesis if the p-value is less than or equal to alpha.

In this specific instance, the obtained p-value (0.6492) is significantly larger than the threshold of 0.05. Therefore, we **fail to reject the null hypothesis** (\$H_0\$).

The statistical evidence gathered from the sample of 500 voters is insufficient to conclude that there is a significant association between gender and political party preference. In practical terms, this result suggests that, based on this survey data, an individual's gender does not statistically predict their political party preference; the two variables appear to be independent.

Summary of Key Procedures in SAS

Successfully executing the Chi-Square Test of Independence in SAS relies on mastering the proper usage of the PROC FREQ procedure, especially when dealing with aggregated frequency data. Remember that using the WEIGHT statement is critical when the dataset contains counts rather than raw observations.

For researchers seeking further depth, the Chi-Square output in SAS also provides other valuable statistics, such as the Likelihood Ratio Chi-Square and Fisher's Exact Test (particularly useful for tables with small expected cell counts). Always ensure your variables are truly categorical variables before running this specific test.