How to Perform a Chi-Square Test of Independence in Stata: A Step-by-Step Guide

Authored by

stats writer

March 9, 2026

Understanding the Chi-Square Test of Independence

The **Chi-Square Test of Independence** is a fundamental **statistical analysis** used to evaluate whether there is a significant association between two **categorical variables**. Unlike correlation coefficients that measure the strength of a linear relationship between continuous data, this test focuses on the distribution of frequencies across different categories. By comparing the observed frequencies in a **contingency table** to the frequencies we would expect if the variables were completely independent, researchers can determine if the patterns they see are due to chance or a meaningful relationship.

In the context of **Stata**, performing this test is both efficient and highly customizable. The software allows users to organize raw data into cross-tabulations, providing a clear visual representation of how categories overlap. This specific **hypothesis test** is widely utilized across various fields, including sociology, medicine, and market research, where data is often grouped into distinct classifications such as gender, region, or success/failure outcomes. Understanding the underlying logic of the test is crucial before executing the commands in a software environment.

To ensure the validity of the results, it is important to verify that the data meets certain **statistical assumptions**. For instance, the observations must be independent, and the expected frequency in each cell of the contingency table should generally be five or greater. If these conditions are met, the **Pearson Chi-Square** statistic provides a reliable measure of the discrepancy between observed and expected counts. In this guide, we will walk through the practical application of this test using one of **Stata**'s most iconic datasets, ensuring you have a comprehensive understanding of the process from data loading to final interpretation.

Theoretical Foundations and Statistical Hypotheses

Before diving into the **Stata** syntax, one must understand the **null hypothesis** (H?) and the **alternative hypothesis** (H?) associated with the test. The null hypothesis states that the two variables are independent, meaning that knowing the value of one variable does not provide information about the value of the other. Conversely, the alternative hypothesis suggests that there is a significant association or dependency between the variables. **Statistical significance** is typically determined by comparing the calculated **p-value** against a predetermined alpha level, most commonly 0.05.

The mathematical core of the **Chi-Square Test of Independence** involves calculating the difference between observed and expected frequencies for each cell. The expected frequency for any given cell is calculated by multiplying the row total by the column total and then dividing by the grand total. These differences are squared, divided by the expected values, and summed to produce the **Chi-Square statistic**. A larger statistic indicates a greater divergence from what

would be expected under the null hypothesis, potentially leading to its rejection.

In **Stata**, these complex calculations are automated, but the researcher remains responsible for the conceptual interpretation. It is essential to remember that a significant result does not imply **causality**; it only indicates that an association exists. For example, finding an association between education level and voting behavior does not necessarily mean education causes specific voting choices, but rather that these variables are linked in the population being studied. This distinction is vital for accurate data storytelling and scientific reporting.

Introduction to the Stata Environment and Dataset

This tutorial utilizes the **auto dataset**, a classic sample dataset built into **Stata** that contains information on 74 different automobiles from the year 1978. This dataset is excellent for learning because it contains a mix of continuous and **categorical variables**. For our analysis, we will focus on two specific attributes that help us explore the relationship between vehicle origin and maintenance history. By using built-in data, you can easily follow along without needing to import external files.

The variables of interest for this specific **Chi-Square Test of Independence** are:

rep78: This variable represents the **repair record** of the vehicle in 1978, coded on a scale from 1 to 5, where 1 indicates a poor record and 5 indicates an excellent record.
foreign: This is a **binary variable** indicating the manufacturing origin of the car, where 0 represents a domestic (US-made) vehicle and 1 represents a foreign-made vehicle.

Our goal is to determine if the location of manufacture (foreign vs. domestic) has a statistically significant relationship with the frequency of repairs the car received. This involves checking if foreign cars were more or less likely to fall into specific repair categories compared to their domestic counterparts. This practical example provides a clear roadmap for applying the **Pearson Chi-Square** methodology to real-world scenarios.

Step 1: Loading and Inspecting the Raw Data

The first stage in any **data analysis** project in **Stata** is to load the data into the system memory. To access the built-in automobile data, we use the **sysuse** command. This command is specifically designed to call up example datasets stored within the **Stata** directory. Type the following command in your **command window**:

```
sysuse auto
```

Once the dataset is loaded, it is a best practice to inspect the raw data to understand its structure

and identify any potential issues, such as **missing values**. We can open the Data Editor in browse mode to view the observations without accidentally changing any values. This is done using the **browse** command, often abbreviated as **br**:

br

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15	Domestic
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179	43	140	3.08	Domestic
26	Linc. Continental	11,497	12	3	3.5	22	4,840	233	51	400	2.47	Domestic
27	Linc. Mark V	13,594	12	3	2.5	18	4,720	230	48	400	2.47	Domestic
28	Linc. Versailles	13,466	14	3	3.5	15	3,830	201	41	302	2.47	Domestic
29	Merc. Bobcat	3,829	22	4	3.0	9	2,580	169	39	140	2.73	Domestic
30	Merc. Cougar	5,379	14	4	3.5	16	4,060	221	48	302	2.75	Domestic
31	Merc. Marquis	6,165	15	3	3.5	23	3,720	212	44	302	2.26	Domestic

As shown in the data browser, each row represents a unique automobile model. You will see columns for various metrics such as price, mileage (mpg), weight, and length. However, for our **Chi-Square Test**, we will ignore the continuous variables and focus exclusively on the **rep78** and **foreign** columns. Observing the data helps confirm that these variables are indeed categorical and suitable for **cross-tabulation** analysis.

Step 2: Performing the Chi-Square Test of Independence

With the data successfully loaded and inspected, we proceed to the core **statistical analysis**. In **Stata**, the **contingency table** and the associated **Chi-Square statistic** are generated simultaneously using the **tabulate** command (often shortened to **tab**). This command is incredibly versatile, allowing for two-way tables that summarize the interaction between two variables.

To perform the **Chi-Square Test of Independence**, you must include the **chi2** option after the comma in your syntax. The general syntax structure is **tab , chi2**. For our specific study, enter the following command into **Stata**:

tab rep78 foreign, chi2

```
. tab rep78 foreign, chi2
```

Repair Record 1978	Car type Domestic	Foreign	Total
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

Pearson chi2(4) = 27.2640 Pr = 0.000

By executing this command, **Stata** produces a **cross-tabulation** displaying the frequency of each category combination. Below the table, you will see the calculated **Pearson Chi-Square** value, the **degrees of freedom**, and the **p-value**. These metrics are the essential components needed to draw a definitive conclusion about the relationship between your chosen variables.

Step 3: Interpreting the Frequency Summary Table

The output begins with a summary table, often called a **joint frequency distribution**. This table is vital for **descriptive statistics**, as it provides a raw count of how many cars fall into each intersection of the two variables. In our example, the rows represent the repair records (1 through 5) and the columns represent the origin (Domestic vs. Foreign).

By examining the table, we can observe several interesting patterns in the **sample** data:

Among domestic cars, only 2 models received the lowest repair rating (1), while 27 models received a middle-of-the-road rating (3).
For domestic vehicles, only 2 models achieved the highest rating of 5, suggesting a potential trend in maintenance performance for that era.
Conversely, foreign cars showed a different distribution, with 9 models receiving a rating of 5 and none receiving a rating of 1 or 2.

These **observed frequencies** provide the raw evidence used to calculate the test statistic. While the table alone suggests that foreign cars might have better repair records, we cannot rely on visual inspection alone. We must look at the **statistical significance** reported below the table to ensure that these differences are not simply the result of **sampling error**.

Step 4: Decoding the Pearson Chi-Square Statistic and P-Value

The bottom section of the **Stata** output contains the results of the **Pearson Chi-Square** test. The first value to note is the **Pearson chi2(4)**, which is 27.2640. The number in parentheses, 4, represents the **degrees of freedom** (df). For a test of independence, the degrees of freedom are calculated as (number of rows - 1) multiplied by (number of columns - 1). In this case, (5-1) * (2-1) = 4.

The most critical value for **statistical inference** is the **Pr** value, which is the **p-value**. In our results, the p-value is 0.000. This indicates the probability of obtaining a **Chi-Square statistic** as extreme as 27.2640 if the **null hypothesis** were true. Because 0.000 is significantly lower than the standard **alpha level** of 0.05, we have strong evidence to reject the null hypothesis.

When we reject the null hypothesis, we conclude that the variables are not independent. In practical terms, this means there is a **statistically significant association** between the manufacturing origin of a car and its repair record. The data suggests that foreign and domestic cars did not have the same distribution of repair frequencies in 1978. This finding allows researchers to move forward with more detailed analyses to explore the nature and direction of this relationship.

Advanced Options and Best Practices in Stata

While the basic **chi2** option is sufficient for many analyses, **Stata** offers additional tools to refine your **Chi-Square Test of Independence**. If you are working with a small **sample size** where some cells have very low expected frequencies, the Pearson Chi-Square may not be accurate. In such cases, you should use the **Fisher's exact test** by adding the **exact** option to your command. This provides a more precise p-value for small or sparse datasets.

Furthermore, you can request **expected frequencies** to be displayed alongside observed frequencies by using the **expected** option. This is helpful for verifying the assumptions of the test. To better understand the contribution of each category, you might also use the **column** or **row** options to see percentages instead of just counts. For example:

tab rep78 foreign, chi2 expected column

This command would show you the percentage of foreign versus domestic cars within each repair

category, making the differences much easier to communicate to a non-technical audience. It is also common to report **effect size** measures, such as Cramer's V, which can be obtained in **Stata** using the **V** option. These measures tell you not just if an association is significant, but how strong that association actually is.

Concluding Thoughts on Chi-Square Analysis

The **Chi-Square Test of Independence** is an essential tool in the **data analyst's** toolkit, providing a clear method for examining relationships between **categorical variables**. By following the steps outlined in this tutorial--loading the data, creating a **contingency table**, and interpreting the **p-value**--you can confidently determine whether observed patterns in your data represent real associations in the population. **Stata**'s intuitive command structure makes this process seamless, allowing you to focus on the implications of your findings.

As you continue your journey with **statistics**, remember that the **Chi-Square test** is just the beginning. It identifies that a relationship exists, but further exploratory data analysis or **regression modeling** may be needed to understand the complexities of that relationship. Always ensure your data meets the necessary **statistical assumptions** to maintain the integrity of your research and produce results that are both reliable and valid.

Finally, clear documentation of your **Stata** code and careful reporting of your **degrees of freedom** and test statistics are vital for **reproducibility**. Whether you are conducting academic research or business analytics, mastering the **Chi-Square Test of Independence** empowers you to extract meaningful insights from categorical data, turning raw numbers into actionable knowledge.