How to Run a Chi-Square Test in Stata to Determine Variable Independence

Authored by

stats writer

December 28, 2025

The Chi-Square Test of Independence is one of the most fundamental statistical tools employed to assess the relationship between two non-numeric, or categorical variables. This powerful technique helps researchers determine whether the distribution of one variable is independent of the distribution of the second variable, or if there is a statistically significant association between them. In the context of the popular statistical software package, Stata, this test is executed using the concise and efficient Stata command "`chi2`".

When running this command in Stata, the input typically mirrors that of the basic tabulation command, `tab`, requiring the specification of the two categorical variables being analyzed. The resulting output is comprehensive, providing critical metrics such as the Chi-Square statistic, the associated p-value, and the crucial measure of degrees of freedom. Understanding these metrics is essential for drawing accurate conclusions regarding the strength and significance of the observed association. Ultimately, the calculated p-value dictates whether we can reject the assumption of independence between the variables.

A **Chi-Square Test of Independence** is specifically formulated to determine whether an observed association between two attributes is statistically significant or merely due to random chance. It operates under a strict set of assumptions, primarily that the variables are categorical (nominal or ordinal) and that the expected cell counts are sufficiently large. Failure to meet these criteria may necessitate the use of alternative tests, such as Fisher's exact test. However, for standard applications, the Chi-Square test provides a robust framework for hypothesis testing involving cross-tabulated data.

This detailed tutorial serves as an essential guide, explaining step-by-step how to efficiently execute and rigorously interpret the results of a Chi-Square Test of Independence within the professional environment of Stata. We will transition from loading a sample dataset to interpreting the final statistical decision, ensuring clarity at every stage of the process.

Establishing the Research Context and Data Selection

To demonstrate the practical application of this statistical procedure, we will leverage a well-known, built-in dataset within the Stata software package. This dataset, conventionally referred to as *auto*, contains detailed information and specifications for 74 distinct automobile models originating from the year 1978. This dataset is frequently utilized in statistical tutorials due to its clear structure and variety of variables, which make it ideal for illustrating bivariate relationships.

Our primary goal using this example is to systematically perform a Chi-Square Test of Independence. The specific research question we aim to address is whether there exists a statistically significant association between two distinct variables contained within the *auto* dataset. In statistical terms, we are testing the null hypothesis that these two variables are entirely

independent against the alternative hypothesis that they are associated.

The two specific categorical variables chosen for this deep dive are defined as follows. Careful attention must be paid to the definition and scaling of these variables, as the Chi-Square test is sensitive to how categories are structured and counted.

rep78: This variable measures the repair record of the vehicle in 1978. It is scaled as an ordinal variable, ranging numerically from 1 (poor repair record) up to 5 (excellent repair record).
foreign: This is a binary, or dichotomous, variable that classifies the origin of the car model. It is coded such that 0 signifies a domestic car type (no, not foreign), and 1 signifies an imported car type (yes, foreign).

Step 1: Loading and Initial Inspection of the Dataset

The foundational step in any data analysis project within Stata involves correctly loading the data into the active memory environment. Since the *auto* dataset is distributed natively with the software, we can access it directly using the system utility command. This ensures reproducibility and ease of access for all users following this tutorial.

To load the data, we execute the following straightforward command in the Stata command window. This action immediately makes the dataset available for manipulation and analysis:

sysuse auto

After successfully loading the dataset, it is considered best practice to perform an initial visual inspection of the raw data structure. This preliminary review, often known as "data snooping," allows the researcher to verify that the data has loaded correctly, confirm the presence of the variables of interest (*rep78* and *foreign*), and identify any potential issues like missing values or unexpected coding before proceeding to the actual statistical test. We achieve this inspection using the standard browsing command:

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 displayed in the data browser snapshot, each row fundamentally represents a single car observation. Accompanying data columns detail various characteristics such as the vehicle's price, miles per gallon (mpg), weight, length, and, critically for our analysis, the values for *rep78* and *foreign*. While the dataset contains a wealth of information, our focus must remain strictly on the two categorical variables essential for conducting the Chi-Square Test.

Step 2: Executing the Chi-Square Test of Independence

Once the data is loaded and verified, the next step is the execution of the primary statistical test. The Stata command structure for performing the Chi-Square Test of Independence is remarkably efficient, integrating directly with the tabulation functionality. By simply appending the `chi2` option to the standard `tabulate` command (or `tab`), Stata is instructed to not only produce the cross-tabulation table but also calculate the necessary statistical figures for the test.

The general syntax required for this operation is intuitive and logical:

tab first_variable second_variable, chi2

Applying this generalized syntax to our specific variables of interest, *rep78* and *foreign*, we arrive at

the exact command required to assess the association between the car's repair record and its origin:

tab rep78 foreign, chi2

Upon execution, Stata generates the comprehensive output matrix that serves as the basis for our statistical conclusion. This output integrates the descriptive count data alongside the inferential statistics necessary to evaluate the null hypothesis of independence.

```
. 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

Interpreting the Cross-Tabulation Summary Table

The first, and highly informative, component of the Stata output is the cross-tabulation table itself, sometimes labeled the **Summary table**. This matrix meticulously displays the frequency distribution, providing the raw counts (observed frequencies) for every possible combination of categories across the two variables, *rep78* and *foreign*. Understanding this descriptive matrix is crucial before moving to the inferential statistics.

The table is structured such that the rows represent the categories of the first variable (*rep78*, the repair record), and the columns represent the categories of the second variable (*foreign*, the origin). The intersection of each row and column provides the count of vehicles that satisfy both conditions simultaneously. For instance, by inspecting the cell counts, we can derive highly specific descriptive statistics about the 74 cars in the sample:

When examining cars with the worst repair record (Code 1), there were 2 cars that were domestic and received 1 repair in 1978.
For cars with a slightly better repair record (Code 2), the count rises to 8 domestic cars, indicating

that this category is slightly more populated.
The most common repair record for domestic cars in this sample appears to be Code 3, with 27 cars recorded in this cell, reflecting a higher frequency of average repair quality among domestic models.

This detailed breakdown continues for all repair categories (1 through 5) across both domestic and foreign car types. Furthermore, the table provides marginal totals--the totals for each row and column--which summarize the overall distribution of each variable independently, providing a complete picture of the sampled data distribution.

Analyzing the Pearson Chi-Square Statistic

Following the descriptive summary, Stata presents the inferential results, beginning with the calculation of the Pearson Chi-Square statistic. This statistic, often simply labeled **Pearson chisq(4)** in the output, is the numerical quantification of the difference between the observed frequencies (the counts in the summary table) and the frequencies that would be expected if the two variables were perfectly independent. The number in parentheses, (4), represents the degrees of freedom (df) associated with this specific test.

The calculation for degrees of freedom in a contingency table is determined by the formula: `df = (R - 1) * (C - 1)`, where R is the number of rows (categories of *rep78*, which is 5) and C is the number of columns (categories of *foreign*, which is 2). Thus, `(5 - 1) * (2 - 1) = 4 * 1 = 4`. This parameter is crucial because it defines the shape of the theoretical Chi-Square distribution used to determine the significance of the calculated test statistic.

In this particular instance, the calculated value for the Chi-Square test statistic is **27.2640**. A larger value for the Chi-Square statistic generally suggests a greater discrepancy between the observed data and what would be expected under the assumption of independence. However, the magnitude alone is insufficient to draw a definitive conclusion; we must compare this value to the critical value of the distribution, or, more commonly in modern statistical practice, rely on the associated p-value.

The Critical Role of the P-Value in Decision Making

The subsequent line of output, labeled **Pr**, presents the p-value. This metric represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated (27.2640), assuming that the null hypothesis (that the variables are independent) is true. The p-value is the cornerstone of hypothesis testing, allowing researchers to quantify the evidence against the null hypothesis.

In our output, the p-value is reported as **0.000**. When interpreting this result, we must compare it

against a predetermined significance level, often denoted as alpha (α), which is conventionally set at 0.05. The decision rule is straightforward: if the calculated p-value is less than the significance level ($p < \alpha$), we reject the null hypothesis.

Since our calculated p-value of 0.000 is demonstrably smaller than the standard threshold of 0.05, we have reached a critical statistical conclusion. We must formally reject the null hypothesis that the car's repair record (*rep78*) and its origin (*foreign*) are independent variables.

Drawing Definitive Conclusions and Statistical Reporting

The rejection of the null hypothesis leads directly to the acceptance of the alternative hypothesis: there is compelling and sufficient statistical evidence to conclude that a statistically significant association exists between whether or not a car is foreign and the total number of repairs it received in 1978. In practical terms, the origin of the car is systematically related to its repair record, meaning that the two variables are dependent.

It is important to note that while the Chi-Square Test of Independence confirms the existence of an association, it does not quantify the strength or direction of that relationship. To understand the practical significance--how strongly related the variables are--a researcher might subsequently employ measures of association like Cramer's V or Phi, which are often provided by Stata or can be calculated using additional options.

In reporting these findings, a researcher would typically state: "A Chi-Square Test of Independence revealed a significant association between vehicle origin (Domestic vs. Foreign) and repair record, $\chi2(4) = 27.26$, $p < 0.001$. This finding suggests that repair frequency is dependent on whether the car is foreign or domestic." This complete reporting ensures all necessary statistical parameters--the test statistic, the degrees of freedom, and the p-value--are provided for transparency and review.