Chi-Square Test of Independence with Stata: A Tutorial for Analyzing Categorical Data

Authored by

Mohammed looti

November 8, 2025

The Chi-Square Test of Independence is a foundational tool in inferential statistics, widely applied across fields from social research to medical epidemiology. Its primary purpose is to determine whether there is a statistically significant association between two factors, both of which are measured as categorical variables. When researchers classify data into discrete, non-overlapping groups--such as demographic categories, regions, or satisfaction levels--this test provides a robust framework for evaluating hypotheses concerning the joint distribution of these factors.

Developing proficiency in this technique is essential for sophisticated quantitative analysis. This detailed guide offers a thorough examination of the theoretical principles underlying the test and provides precise, step-by-step instructions on how to execute a **Chi-Square Test of Independence** efficiently using the powerful statistical software package, Stata. By following these steps, analysts can move confidently from data loading to formal statistical conclusion.

Understanding the Theoretical Basis of the Chi-Square Test

At its core, the **Chi-Square Test of Independence** functions by comparing the actual counts observed in a dataset (the observed frequencies) against the counts that would theoretically be expected if the two variables were completely independent (the expected frequencies). The concept of independence means that the distribution of one variable is not influenced or affected by the categories of the other. The resulting Chi-Square test statistic quantifies the total cumulative discrepancy between these observed and expected values; a larger difference indicates stronger evidence of an association.

The formal structure of hypothesis testing for this procedure relies on two competing statements. The **Null Hypothesis** (H_0) asserts that there is absolutely no relationship or association between the two categorical factors--they are independent. Conversely, the Alternative Hypothesis (H_a) posits that a statistically significant association exists between the variables, meaning they are dependent. A high calculated value for the test statistic provides compelling empirical grounds to reject the null hypothesis in favor of the alternative.

It is crucial to ensure that the assumptions necessary for the Chi-Square test are met to guarantee the validity of the results. The key preconditions include that the data must be collected through random sampling, and the variables under scrutiny must be either nominal or ordinal categorical variables. Furthermore, a practical rule of thumb suggests that the expected frequencies for the majority of cells within the contingency table should ideally be greater than five. This condition ensures that the calculated Chi-Square distribution serves as an accurate approximation for the true sampling distribution of the test statistic.

Setting Up the Analysis: Dataset and Research Question

To provide a clear, practical demonstration of the procedure within the Stata environment, we will

utilize the universally accessible, built-in dataset named *auto*. This dataset compiles extensive information on 74 distinct automobile models manufactured in 1978, detailing various characteristics such as pricing, fuel efficiency (MPG), vehicle weight, and repair history. Our specific analytical objective is to investigate whether a meaningful relationship exists between the vehicle's country of origin and its recorded repair frequency.

The central research question guiding our analysis is: Is there a statistically significant association between a car's origin (defined as domestic or foreign) and the frequency of repairs it required in 1978? This question necessitates the analysis of the joint distribution of two specific variables contained within the *auto* dataset, which are defined below:

rep78: This variable documents the number of times the car received a repair throughout 1978. It is an ordinal variable spanning five discrete categories, coded from 1 (worst) through 5 (best).
foreign: This is a simple binary nominal variable classifying the car's origin, where the value 0 represents a domestic model and the value 1 represents a foreign model.

By focusing the analysis on these two variables, the **Chi-Square Test of Independence** will enable us to formally determine if the distribution of repair records (*rep78*) is statistically homogeneous for both domestic and foreign cars, or conversely, if these characteristics are statistically dependent. The subsequent steps detail the precise execution required to conduct this test within Stata.

Step 1: Data Preparation and Exploration in Stata

The initial procedural requirement in any Stata session is ensuring that the necessary dataset is correctly loaded into memory. Since the *auto* dataset is a standard sample file pre-installed with the software, we can load it instantly using the dedicated `sysuse` command. This command efficiently retrieves the specified data file directly from Stata's system directories, making it immediately available for analysis and manipulation.

To load the required data, execute the following command exactly as shown in the Stata Command Window:

sysuse auto

Once the data is successfully loaded, it is standard practice and highly recommended to perform a preliminary visual inspection of the raw data structure. This exploration confirms the presence of the variables of interest (*rep78* and *foreign*) and verifies that they are structured as anticipated. We can achieve this visualization quickly using the `br` (browse) command, which opens a dynamic, spreadsheet-like window displaying every observation in the dataset.

Input the following command to open the data browser view:

br

The resulting display confirms the structure of the data, where each row corresponds to a single car observation. While the dataset encompasses numerous attributes--including cost, mileage (mpg), and physical dimensions--our statistical focus remains strictly limited to examining the joint relationship between *rep78* and *foreign*. The image below provides a visual representation of the typical raw data view generated upon executing the browse command.

	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

Step 2: Executing the Chi-Square Test Command

With the data successfully loaded and its structure verified, the critical next phase is the execution of the statistical test itself. In Stata, the **Chi-Square Test of Independence** is performed using the versatile `tabulate` command (often abbreviated to `tab`). This command is specifically designed for generating two-way frequency tables, commonly known as contingency tables. To obtain the required Chi-Square statistic, we must explicitly append the `chi2` option to the command syntax.

The conventional syntax for running this associative analysis is highly intuitive and requires only the specification of the two categorical variables whose relationship we intend to test. By

convention, the variable intended for the rows of the table is listed first, followed by the variable designated for the columns, although the statistical conclusion remains symmetrical regardless of this positional arrangement:

tab first_variable second_variable, chi2

Applying this standard template to our specific research question, which involves comparing car repair records (*rep78*) against origin (*foreign*), the exact command necessary for the analysis is as follows. The specified order determines the visual layout of the resulting contingency table.

tab rep78 foreign, chi2

Upon execution, Stata rapidly generates a comprehensive output window. This output simultaneously presents the fully formed contingency table alongside the essential summary statistics required for making a formal statistical decision regarding the Null Hypothesis. This comprehensive result forms the analytical foundation for interpreting whether the car's repair history and its origin are independent or associated factors.

Interpreting the Output: The Contingency Table

The initial and most visually striking component of the Stata output is the **Summary Table**, frequently termed the cross-tabulation or contingency table. This table is foundational because it displays the observed frequencies--the actual counts of vehicles that fall into every possible combination of categories for *rep78* and *foreign*. A careful analysis of this table offers immediate, descriptive insight into the sample's distribution before engaging with the formal inferential statistics.

The table is structured with the categories of *rep78* (Repair Record) arrayed across the rows (categories 1 through 5) and the categories of *foreign* (Car Origin) positioned across the columns (Domestic and Foreign). The intersection point of any row and column provides the count for that specific pairing. For example, by examining the cell counts, we can observe the following raw data distributions:

Two cars classified as **domestic** received a repair rating of 1 in 1978.
Eight cars classified as **domestic** received a repair rating of 2 in 1978.
Twenty-seven cars classified as **domestic** received a repair rating of 3 in 1978.

Visual inspection of these raw counts allows us to formulate preliminary observations regarding the distribution patterns. For instance, domestic cars appear to concentrate heavily in the middle repair categories (3 and 4), whereas foreign cars seem to have a disproportionately higher count in the excellent repair category (5). However, descriptive inspection alone is statistically insufficient; we

must rely on the formal results of the statistical test to determine if these observed sample differences reflect true population dependence or are merely attributable to random sampling chance.

```
. 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 Output: Statistical Significance and Conclusion

The most crucial section of the output is located immediately beneath the contingency table, where Stata reports the calculated test statistic and its corresponding p-value. These two values are paramount as they allow us to formally reject or fail to reject the Null Hypothesis of independence.

The line labeled **Pearson chisq(4)** reports the specific value calculated for the Chi-Square test statistic, which is 27.2640 in this output. The number enclosed in parentheses, (4), represents the degrees of freedom (df) associated with the test. The degrees of freedom are mathematically derived from the dimensions of the contingency table using the formula: $(R-1) times (C-1)$, where R is the number of rows (5 repair categories) and C is the number of columns (2 origin categories), yielding 4 degrees of freedom. This value is fundamentally important for defining the shape of the theoretical Chi-Square distribution used to calculate the probability of observing our statistic.

The final crucial element is the row labeled **Pr**, which displays the associated probability value, or p-value. In this specific analysis, the reported value is 0.000 (indicating that the p-value is less than 0.0005). The established decision rule for formal hypothesis testing dictates that if the p-value falls below a predefined significance level (alpha, conventionally set at 0.05), we must reject the Null Hypothesis. Since 0.000 is substantially less than the 0.05 threshold, we confidently reject the hypothesis that the two variables are independent.

In conclusion, the statistically significant result provides strong, quantitative evidence to confirm

that there is a meaningful and non-random association between a car's country of manufacture (domestic or foreign) and the total number of repairs it required in 1978. This analysis demonstrates that the distribution of repair frequency is significantly different across the car origin categories, establishing that origin and repair history are indeed dependent variables.