Chi-Square Goodness of Fit Test in Stata: A Step-by-Step Guide

Authored by

Mohammed looti

November 8, 2025

The **Chi-Square Goodness of Fit Test** represents a fundamental and indispensable statistical procedure utilized across various empirical disciplines, ranging from social sciences to bioinformatics. Its primary function is to rigorously assess whether the observed distribution of frequencies for a specific categorical variable within a collected sample deviates significantly from a theoretical, predetermined, or previously established distribution. This comparison is critical for validating assumptions about population characteristics or testing hypotheses derived from theoretical models.

This powerful test provides researchers with the necessary framework to compare actual sample proportions against known or expected population proportions, moving beyond simple descriptive statistics to robust inferential analysis. Understanding when and how to apply this test is a cornerstone of quantitative research methodology, ensuring that conclusions drawn about a population based on sample data are statistically sound and reliable.

This comprehensive tutorial is meticulously designed to guide the user through the entire process of executing and interpreting the Chi-Square Goodness of Fit Test using **Stata**, one of the most widely respected and powerful statistical software packages available today. We will detail the essential steps required, including data preparation, installation of specialized user-written commands, execution of the core analysis, and a thorough explanation of how to derive statistically meaningful conclusions from the generated output. Our focus is on achieving analytical rigor and reproducibility.

Theoretical Foundations of the Goodness-of-Fit Test

Before commencing the practical application within Stata, grasping the theoretical underpinnings of this statistical test is absolutely essential. The Goodness of Fit Test operates fundamentally by contrasting the set of frequencies actually observed in the collected dataset—the sample data—with the set of frequencies that would be expected if the underlying assumptions, defined by the **null hypothesis**, were perfectly true.

The **null hypothesis** (H0) in this context always posits that there is no statistically significant difference between the observed frequency distribution and the hypothesized distribution. Conversely, the alternative hypothesis (Ha) asserts that a statistically significant difference does exist, meaning the sample distribution does not conform to the expected population proportions. The core objective of the test is to determine if the discrepancies observed are large enough to warrant rejecting the null hypothesis in favor of the alternative.

The calculation of the test statistic, known as the **Chi-Square test statistic**, is based on the summation of the squared differences between the observed and expected frequencies, weighted by the expected frequencies themselves. Mathematically, a large resulting Chi-Square value signifies a substantial divergence between what was collected empirically and what was

hypothesized theoretically. This substantial discrepancy suggests that the fit between the observed data and the hypothesized model is poor, thereby leading toward the rejection of H0. Conversely, a small Chi-Square value indicates a close alignment, suggesting the null hypothesis should be retained.

A crucial aspect governing the validity and reliability of the Chi-Square approximation is the satisfaction of specific assumptions, particularly regarding sample size. For the test results to be robust, it is generally required that the **expected frequency** in every single category must be sufficiently large—standard guidelines typically recommend that all expected frequencies should be five or greater. Failure to meet this assumption may necessitate combining categories or employing alternative statistical procedures. This methodology is vital for testing claims across diverse applications, such as evaluating whether market share is equally distributed among competitors or, as demonstrated in our forthcoming analysis, assessing if sample demographics accurately mirror known population proportions.

Example: Assessing Race Distribution in Labor Statistics

To provide a clear, practical demonstration of implementing the Chi-Square Goodness of Fit Test, we will utilize a well-known, built-in **Stata** dataset, *nlsw88*. This dataset is a valuable resource containing detailed labor statistics gathered in 1988 for a cohort of young women in the United States, encompassing variables related to employment, education, and detailed demographics. Our specific focus for this analysis centers on the *race* variable, which is a key categorical variable.

Our primary objective is to subject the distribution of the *race* variable within this sample to a rigorous Goodness of Fit Test against a specific, predefined population hypothesis. For the purpose of this illustration, we hypothesize that the true distribution of race within this sampled population adheres precisely to the following proportions: **70% White, 20% Black, and 10% Other**. This hypothetical distribution serves as our baseline, the foundation of the **null hypothesis** (H0) we seek to test.

The resulting analysis will scientifically determine whether the observed proportions of race categories in the *nlsw88* sample exhibit a statistically significant deviation from these expected population percentages. If the deviation is significant, it implies the sample is not representative of the hypothesized population distribution. We will execute this analysis through a meticulously structured series of steps, beginning with data preparation and culminating in the command execution and a detailed, statistically sound interpretation of the results generated by Stata.

Step 1: Data Acquisition and Initial Review

The initial and paramount step in any Stata analysis is ensuring the relevant data is correctly loaded and accessible within the environment. Since *nlsw88* is a widely recognized and standard

dataset included with the Stata installation package, we can load it efficiently using the system utility command, circumventing the need for external file paths.

To load the dataset into the current Stata session, the analyst must execute the following simple command directly in the Stata Command window:

sysuse nlsw88

Upon successful loading, it is considered best practice and essential for data quality assurance to conduct a preliminary visual inspection of the dataset. This step allows the researcher to quickly confirm the variable names, verify data types (ensuring *race* is indeed coded categorically), and gain an overall sense of the data structure and scope. This critical review ensures that the intended variable is prepared correctly for the inferential statistical procedure that follows.

To facilitate this visual inspection and view the raw data observations within the Stata Data Browser, execute the following concise command:

br

	idcode	age	race	married	never_marr~d	grade	collgrad	south	smsa	c_city	industry
1	1	37	black	single	0	12	not college grad	0	SMSA	0	Transport/Comm/Utility
2	2	37	black	single	0	12	not college grad	0	SMSA	1	Manufacturing
3	3	42	black	single	1	12	not college grad	0	SMSA	1	Manufacturing
4	4	43	white	married	0	17	college grad	0	SMSA	0	Professional Services
5	6	42	white	married	0	12	not college grad	0	SMSA	0	Manufacturing
6	7	39	white	married	0	12	not college grad	0	SMSA	0	Professional Services
7	9	37	white	single	0	12	not college grad	0	SMSA	1	Transport/Comm/Utility
8	12	40	white	married	0	18	college grad	0	SMSA	0	Professional Services
9	13	40	white	married	0	14	not college grad	0	SMSA	0	Professional Services
10	14	40	white	married	0	15	not college grad	0	SMSA	0	Professional Services
11	15	39	white	married	0	16	college grad	0	SMSA	0	Professional Services
12	16	40	white	married	0	15	not college grad	0	SMSA	0	Professional Services
13	18	40	white	married	0	15	not college grad	0	SMSA	0	Wholesale/Retail Trade
14	19	40	white	single	0	15	not college grad	0	SMSA	0	Professional Services
15	20	39	white	married	0	15	not college grad	0	SMSA	0	Professional Services
16	22	41	white	married	0	15	not college grad	0	SMSA	0	Professional Services
17	23	42	white	married	0	15	college grad	0	nonSMSA	0	Professional Services
18	24	41	white	married	0	14	college grad	0	SMSA	0	Professional Services
19	25	42	white	married	0	14	college grad	0	SMSA	1	Professional Services
20	36	37	white	single	1	12	not college grad	0	SMSA	0	Business/Repair Svc
21	39	44	white	single	0	16	college grad	0	SMSA	0	Professional Services
22	44	41	white	married	0	18	college grad	0	SMSA	0	Public Administration
23	45	35	white	married	0	12	not college grad	0	SMSA	0	Transport/Comm/Utility
24	46	44	white	married	0	18	college grad	0	SMSA	0	Professional Services
25	47	35	white	single	0	12	not college grad	0	SMSA	0	Transport/Comm/Utility
26	48	35	white	single	0	15	not college grad	0	SMSA	1	Finance/Ins/Real Estate
27	50	36	white	single	0	16	college grad	0	SMSA	1	Professional Services
28	51	38	white	married	0	12	not college grad	0	SMSA	0	Professional Services
29	54	40	white	single	1	12	not college grad	0	SMSA	1	Professional Services
30	57	42	white	married	0	12	not college grad	0	SMSA	1	Wholesale/Retail Trade
31	62	38	white	married	0	10	not college grad	0	nonSMSA	0	Wholesale/Retail Trade
32	63	44	white	single	0	15	college grad	0	SMSA	1	Professional Services
33	64	38	white	married	0	12	not college grad	0	SMSA	0	Professional Services

The Data Browser provides a comprehensive table where each row corresponds to an individual observation (a specific woman in the survey), displaying attributes such as age, employment status, educational level, and, critically for this analysis, her recorded race category. This preliminary step confirms that the dataset contains a total of 2,246 unique observations. This total sample size is fundamental, as it forms the basis upon which all subsequent calculations of expected frequencies will be derived, thereby linking our hypothesized percentages directly to the actual size of the sample under investigation.

Step 2: Installing Specialized Goodness of Fit Command

While Stata boasts an extensive library of built-in statistical commands, the specific functionality required to efficiently perform the Chi-Square Goodness of Fit Test when testing against custom, user-defined expected percentages is provided by a community-developed, user-written package. This package is named *csgof*, and its use exemplifies the exceptional flexibility and extensibility offered by the Stata user community, which continually develops specialized tools tailored for specific analytical demands.

Before the primary analysis can be executed, this essential package must be located and correctly installed within the Stata environment. To initiate the search for the *csgof* package, the following command must be entered:

findit csgof

Executing the `findit` command will prompt Stata to open a new viewer window, which lists all potentially relevant resources, including official help files and, crucially, user-contributed packages. The user must carefully navigate this list and select the link that corresponds to the package installation source, which is typically hosted on authoritative statistical or academic websites, often labeled similarly to *csgof from https://stats.idre.ucla.edu/stat/stata/ado/analysis*. This ensures the acquisition of a reliable version of the command.

Following the selection of the correct link, a dedicated installation window will appear. The user should then click the prompt that explicitly instructs them to *click here to install*. Stata will then automatically manage the entire download and installation process, making the *csgof* command immediately available for utilization throughout the current session. Successful completion of this installation step is a non-negotiable prerequisite for proceeding to the execution phase in Step 3, and users should ensure they possess adequate network connectivity and system permissions to avoid installation failures.

Step 3: Executing the Goodness-of-Fit Test

With the *csgof* package successfully installed and the data loaded, we are prepared to execute the

focal point of our analysis: the Chi-Square Goodness of Fit Test. As previously established, our aim is to rigorously test whether the observed distribution of the *race* variable deviates significantly from our hypothesized population distribution defined by the percentages: 70% White, 20% Black, and 10% Other.

The structure required for the *csgof* command is designed for clarity and precision, necessitating only the input of the primary categorical variable being analyzed and the full list of expected percentages for each category. The general syntax adheres strictly to the following format:

csgof variable_of_interest, expperc(list_of_expected_percentages)

It is absolutely critical to ensure that the sequential order of the expected percentages provided within the `expperc()` option precisely matches the internal ordering of the categories as defined by Stata for the variable. For the *race* variable in the *nlsw88* dataset, the categories are ordered numerically as 1 (White), 2 (Black), and 3 (Other). Therefore, to accurately reflect our 70%, 20%, and 10% hypothesis, the exact command syntax that must be executed is:

csgof race, expperc(70, 20, 10)

```
. csgof race, expperc(70, 20, 10)
```

race	expperc	expfreq	obsfreq
white	70	1572.2	1,637
black	20	449.2	583
other	10	224.6	26

```
chisq(2) is 218.13, p = 0
```

Step 4: Comprehensive Interpretation of Stata Output

The statistical output generated by the *csgof* command is divided into two distinct but interconnected sections, both requiring meticulous examination: the detailed summary box and the

inferential statistics reported at the foot of the output. The summary box serves as a direct, empirical comparison between the hypothetical distribution defined by the researcher and the actual distribution observed within the sample data.

The summary box offers a critical, side-by-side juxtaposition of the observed and expected characteristics for each distinct category of the *race* variable, providing immediate insight into the nature and magnitude of the discrepancies:

Expected Percent: This column serves as a confirmation of the percentages that were inputted into the `expperc()` option (70%, 20%, 10%). It represents the precise theoretical distribution against which the sample data is being tested.
Expected Frequency: This value calculates the absolute number of observations (individuals) that we would statistically anticipate finding in each race category, assuming the **null hypothesis** of perfect fit were absolutely true. Given the total sample size of 2,246, the expected count for White individuals is calculated as 70% of 2,246, resulting in 1,572.2. Correspondingly, the Black category is expected to contain 449.2 individuals (20%), and the Other category, 224.6 individuals (10%). These values are crucial for fulfilling the minimum expected frequency assumption.
Observed Frequency: This column displays the actual, empirical count of individuals recorded in the *nlsw88* dataset for each corresponding category. The observed frequency for White individuals is 1,637; for Black individuals, it is 415; and for Other individuals, it is 194. A preliminary visual inspection immediately highlights the numerical disparity between the Observed and Expected Frequencies across all categories, discrepancies which fundamentally drive the calculation of the Chi-Square test statistic.

The second, and arguably most important, section of the output contains the inferential statistics necessary for making a formal decision regarding the null hypothesis, adhering to the principles of hypothesis testing:

Chisq(2): This reported value, 218.13, is the calculated **Chi-Square test statistic** itself. This statistic quantifies the overall magnitude of the difference between the observed and expected data. The number enclosed in parentheses, (2), denotes the degrees of freedom (df) for the test. In the context of a Chi-Square Goodness of Fit test, the degrees of freedom are always calculated as the number of categories (k) minus one (k - 1). Since our variable has three race categories, the calculation is 3 - 1 = 2 degrees of freedom.
p: This numerical value represents the p-value associated with the calculated Chi-Square test statistic of 218.13. The output reports the p-value as 0.0000 (which should be interpreted as less than 0.0001). The p-value is defined as the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from our sample data, assuming the **null hypothesis** is entirely true. A very low p-value suggests the observed data is highly unlikely under the null model.

Step 5: Decision Making and Final Conclusions

The definitive final step in the hypothesis testing procedure involves utilizing the calculated p-value to evaluate the statistical significance of the results against a predetermined threshold. This threshold, denoted as the significance level (alpha, $alpha$), is conventionally set at 0.05 (or 5%). The fundamental decision rule guiding this assessment is straightforward and universal across parametric and non-parametric tests: if the p-value is less than the chosen alpha level ($p < $alpha$), we are statistically compelled to **reject the null hypothesis**.

In the context of this specific analysis, the calculated p-value is reported as 0.0000. Since this value is substantially and unequivocally less than the predetermined significance level of $alpha = 0.05$, the statistical imperative is clear: we must **reject the null hypothesis** (H0). Recall that H0 explicitly stated that the true distribution of race within the *nlsw88* dataset is consistent with the hypothesized proportions of 70% White, 20% Black, and 10% Other.

Consequently, the formal statistical conclusion is that there exists overwhelming evidence to confidently assert that the observed distribution of the *race* variable within the sample is significantly different from the hypothesized population distribution. The extraordinarily large **Chi-Square test statistic** (218.13) confirms that the magnitude of the discrepancies detected between the observed sample counts and the expected theoretical counts are far too extensive to be reasonably explained by mere random sampling variability or chance. This outcome definitively indicates that the racial composition of the *nlsw88* sample does not align with the hypothetical population percentages initially proposed, suggesting either that the sample is not representative of that specific hypothesized population or that the underlying population proportions are different from the ones proposed.

Mastering the Chi-Square Goodness of Fit Test in Stata empowers researchers to draw rigorous conclusions about population distributions, moving confidently from descriptive summaries to powerful inferential statements based on observed empirical data.