

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

Authored by  
**stats writer**

March 8, 2026

## RECOMMENDED CITATION

stats writer (2026). *How to Perform a Chi-Square Goodness of Fit Test in Stata: A Step-by-Step Guide*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=134681>

## Understanding the Chi-Square Goodness of Fit Test

The **Chi-Square Goodness of Fit Test** serves as a foundational **statistical hypothesis test** designed to determine how well an observed set of data aligns with a specific, predetermined distribution. In the realm of **quantitative research**, this test is indispensable for investigators who need to validate whether their empirical findings deviate significantly from what would be expected under a theoretical model. By comparing the **observed frequencies** within various categories to the **expected frequencies** calculated from a hypothesized distribution, researchers can make informed decisions regarding the validity of their underlying assumptions.

This particular **nonparametric test** is most frequently applied to **categorical variables**, where data points are sorted into mutually exclusive bins or groups. Whether a scientist is examining the phenotypic ratios in a genetic experiment or a sociologist is analyzing the demographic breakdown of a survey, the **Goodness of Fit** analysis provides a standardized metric--the **Chi-Square statistic**--to quantify the discrepancy between reality and theory. A high level of agreement between these values suggests that the model is an accurate representation of the population, whereas a large discrepancy may indicate that the hypothesized distribution is incorrect.

Within the **Stata** software environment, performing this analysis is a streamlined process that involves specific commands and diagnostic tools. **Stata** is a powerful **statistical software** package widely utilized in fields such as **economics**, **epidemiology**, and **political science** due to its robust data management capabilities and extensive library of analytical procedures. Mastering the implementation of the **Chi-Square Goodness of Fit Test** in **Stata** allows researchers to transition from raw data to sophisticated statistical inference with high precision and reproducibility.

## The Significance of Categorical Data Analysis

Before diving into the procedural steps in **Stata**, it is vital to recognize why **categorical variables** require specialized testing like the **Chi-Square**. Unlike **continuous variables**, which can be analyzed using means and standard deviations, categorical data relies on counts and percentages. The **Goodness of Fit test** evaluates the **probability distribution** of these counts across different levels, such as race, gender, or preference rankings. This is critical when researchers have a "null" expectation--for example, expecting an equal distribution across all groups or following a specific ratio derived from historical census data.

The **null hypothesis** for this test typically states that there is no significant difference between the observed counts and the expected counts. In other words, any observed variation is assumed to be the result of **sampling error** rather than a true departure from the hypothesized model. Conversely, the **alternative hypothesis** posits that the data does not follow the specified distribution. This framework allows researchers to rigorously test theoretical claims against real-

world evidence gathered through observation or experimentation.

In many practical scenarios, such as **market research** or public health studies, understanding the distribution of a population is the first step toward more complex **multivariate analysis**. If the sample distribution does not fit the expected population parameters, it may indicate a bias in the **sampling method** or a fundamental shift in the population's characteristics over time. Therefore, the **Chi-Square Goodness of Fit Test** acts as a gatekeeper for data quality and theoretical consistency.

## Preparing the Stata Environment and Loading Datasets

To begin our practical demonstration, we will utilize one of **Stata's** built-in datasets to ensure that the results are reproducible. The **sysuse** command is a convenient way to access example data provided by the developers. For this tutorial, we will focus on the **nlsw88** dataset, which contains **longitudinal survey** data for women in the United States from 1988. This dataset is rich with demographic and economic variables, making it an ideal candidate for testing categorical distributions like race or marital status.

The first step in any **Stata** workflow is to load the data and conduct a preliminary review to understand the **data structure**. By viewing the raw data, you can identify how variables are coded and ensure there are no glaring errors or missing values that might skew the results of your **data analysis**. This initial inspection is a hallmark of good **data hygiene** and sets the stage for a successful statistical test.

### Step 1: Load and view the raw data.

To load the 1988 National Longitudinal Survey data, execute the following command in the **Stata** command window:

```
sysuse nlsw88
```

After loading the data, you can open the Data Browser to inspect the individual observations by typing:

```
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 resulting spreadsheet view displays information for each individual in the sample, including **demographic indicators** such as age, race, and education. Each row represents a unique respondent, while the columns correspond to different variables measured during the survey period.

## Installing External Packages for Advanced Analytics

While **Stata** includes a vast array of native commands, some specialized tasks require the installation of user-written packages from **SSC (Statistical Software Components)** or other repositories. The **Chi-Square Goodness of Fit Test** is traditionally performed using the **csgof** command, which is not part of the base **Stata** installation. This command provides a much more intuitive output specifically tailored for **Goodness of Fit** compared to the standard tabulation commands.

### Step 2: Load the goodness of fit package.

To integrate this functionality into your **Stata** environment, you must use the **findit** command to locate and download the necessary files. This process demonstrates the flexibility of **Stata**,

allowing the community to expand the software's capabilities through **open-source** contributions. Follow the instructions below to install the **csgof** utility:

### **findit csgof**

Upon running this command, a new viewer window will appear within the **Stata** interface. To complete the installation, follow these sub-steps:

Locate and click the hyperlink labeled **csgof** from **<https://stats.idre.ucla.edu/stat/stata/ado/analysis>**.

In the subsequent window, click the link that states "**click here to install**".

Wait for the confirmation in the results window that the files have been successfully installed.

Once the **csgof** package is active, you are ready to perform the analysis. This package automates the calculation of **expected frequencies** based on the percentages you provide, significantly reducing the risk of manual calculation errors.

## **Executing the Statistical Command and Syntax**

With the environment prepared and the data loaded, we can now define our research question. Suppose we hypothesize that the racial distribution in our dataset should mirror a specific theoretical model: 70% White, 20% Black, and 10% Other. The **Chi-Square Goodness of Fit Test** will allow us to determine if the **empirical evidence** in the **nlsw88** dataset supports or refutes this specific distribution.

### **Step 3: Perform the Goodness-of-Fit Test.**

The **csgof** command requires you to specify the **variable of interest** followed by the expected percentages. It is crucial that the percentages you list in the command sum to 100%, and they must be provided in the same order as the levels of the categorical variable (typically in ascending numerical or alphabetical order).

The general syntax for the command is as follows:

**csgof variable\_name, expperc(p1, p2, ..., pn)**

Applying this to our specific hypothesis regarding race, we enter the following command into **Stata**:

**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

This command instructs **Stata** to take the "race" variable, calculate the total number of observations, and then determine how many individuals \*should\* fall into each racial category if the 70/20/10 distribution were perfectly true. It then compares these theoretical values to the actual counts found in the **sample**.

## Analyzing Observed versus Expected Frequencies

The output generated by **Stata** provides a detailed summary table that is essential for **data interpretation**. Understanding each component of this table allows the researcher to see exactly where the data deviates from the model. The output is typically divided into a summary box and the final **test statistics**.

The summary box reveals three critical pieces of information for each category:

**Expected Percent:** This reflects the **hypothesized distribution** you provided. For example, the 70% we assigned to the "White" category appears here as the baseline for comparison.

**Expected Frequency:** This is the raw count that the model predicts. In our dataset of 2,246 individuals, **Stata** calculates that 70% of the total equals 1,572.2. This represents the **mean** count we would expect across many similar samples if the null hypothesis were true.

**Observed Frequency:** This is the actual count of individuals within the "White" category recorded in the survey. Our output shows an observed frequency of 1,637, which is slightly higher than the predicted 1,572.2.

By comparing the **observed values** against the **expected values** across all categories (White, Black, and Other), we can begin to see the "residuals" or the differences that the **Chi-Square** test will aggregate into a single value. If the differences are small, the **Chi-Square statistic** will be low; if the differences are large, the statistic will be high.

## Evaluating Statistical Significance and Hypotheses

The ultimate goal of the test is to look at the **Chi-Square test statistic** and its associated **p-value**. In our specific output, the **Chisq(2)** value is reported as 218.13. The number in parentheses, 2, represents the **degrees of freedom**, which is calculated as the number of categories minus one (3 categories - 1 = 2).

The **p-value** is the most critical metric for determining **statistical significance**. It tells us the probability of obtaining a **Chi-Square statistic** as extreme as 218.13 if the **null hypothesis** were actually true. In our case, the **p-value** is reported as 0 (or less than 0.001). Using a standard **significance level** (alpha) of 0.05, we compare our **p-value** to this threshold.

Because the **p-value** (0.000) is significantly less than 0.05, we **reject the null hypothesis**. This leads to the following conclusions:

There is a **statistically significant** difference between our observed racial distribution and the hypothesized 70/20/10 distribution.

We have sufficient evidence to conclude that the true distribution of race in the 1988 population (as represented by this sample) does not fit the proposed model.

The discrepancy between the 1,637 observed white individuals and the 1,572.2 expected, along with deviations in other categories, is too large to be attributed to random chance alone.

## Methodological Assumptions and Best Practices

While the **Chi-Square Goodness of Fit Test** is a robust tool, its validity depends on several key **statistical assumptions**. First and foremost is the assumption of **independence of observations**. Each subject in the **nlsw88** dataset must represent a unique individual, and the classification of one person should not influence the classification of another. If the data includes repeated measures of the same people, a different test, such as **McNemar's test**, might be more appropriate.

Another critical requirement involves the **sample size** and the size of the **expected frequencies**. A common rule of thumb in **statistics** is that all expected cell counts should be at least 5. If any expected frequency is too small, the **Chi-Square distribution** may not accurately approximate the **sampling distribution** of the test statistic, potentially leading to an inaccurate **p-value**. In our example, the expected frequencies were all well above this threshold (e.g., 1,572.2), ensuring the reliability of our results.

Finally, researchers should always supplement their **p-values** with **effect size** measures or a qualitative review of the residuals. While the test tells us *that* a difference exists, it does not inherently tell us *why* or how meaningful that difference is in a practical context. By reporting both

the **Chi-Square** results and the raw percentage differences, you provide a comprehensive narrative that enhances the transparency and impact of your research findings.

ARABPSYCHOLOGY.COM