How to Test for Normality in Stata: A Step-by-Step Guide

Authored by

stats writer

March 9, 2026

Stata is a sophisticated integrated statistical software package that provides researchers, economists, and data scientists with a robust suite of tools for **data management**, **visualization**, and advanced **econometric modeling**. One of the most fundamental tasks in exploratory data analysis is determining the underlying distribution of a dataset. Specifically, identifying whether a variable follows a **normal distribution** is essential because many classical statistical procedures--including **t-tests**, **ANOVA**, and **linear regression**--rely on the assumption of normality to produce valid **p-values** and **confidence intervals**. When data deviates significantly from this "bell-shaped" curve, the reliability of these parametric tests may be compromised, leading to potentially erroneous conclusions.

In the **Stata** environment, users have access to a diverse array of methodologies to evaluate normality, ranging from intuitive graphical inspections to rigorous **statistical hypothesis testing**. Graphical methods, such as **histograms** and **Q-Q plots**, offer a visual representation of the data's spread, allowing researchers to spot outliers or skewness immediately. On the other hand, formal tests provide a definitive **p-value** that quantifies the evidence against the **null hypothesis** of normality. By combining these approaches, analysts can make informed decisions regarding whether to proceed with parametric methods or to employ **non-parametric** alternatives and **data transformations**.

This comprehensive guide explores the primary techniques available in **Stata** for testing normality. We will examine the practical application of **histograms**, the **Shapiro-Wilk test**, the **Shapiro-Francia test**, and the **Skewness and Kurtosis test**. Each method has specific requirements regarding **sample size** and sensitivity to different types of distributional deviations. Understanding these nuances is critical for any researcher aiming to maintain high standards of **statistical integrity** in their empirical work.

The Theoretical Importance of the Normal Distribution

The **normal distribution**, often referred to as the **Gaussian distribution**, serves as the cornerstone of **inferential statistics**. This importance stems largely from the **Central Limit Theorem**, which posits that the distribution of the sample mean tends toward normality as the sample size increases, regardless of the population's original distribution. However, for smaller samples, the assumption that the data itself is normally distributed remains vital for the validity of **parametric statistics**. If a variable is non-normal, the standard errors of estimates may be biased, which directly impacts the **statistical significance** of the results.

Beyond simple significance testing, normality is a key assumption in the analysis of **residuals** in **regression analysis**. For a linear regression model to be considered the **Best Linear Unbiased Estimator (BLUE)**, the errors should ideally be independent and identically distributed (i.i.d.) following a **normal distribution**. While **Ordinary Least Squares (OLS)** regression is somewhat robust to violations of this assumption in large samples, extreme **skewness** or heavy tails (**kurtosis**) can significantly distort the predictive power and accuracy of the model. Therefore, testing for normality is not merely a box-ticking exercise but a foundational step in ensuring **model specification** is correct.

Researchers must also distinguish between the normality of the raw data and the normality of the error terms. In many instances, independent variables do not need to be normally distributed; however, the dependent variable's relationship with these predictors should result in normally distributed residuals. By mastering the normality tests within **Stata**, analysts can determine if their data requires a **logarithmic transformation**, a **Box-Cox transformation**, or if they should switch to a **generalized linear model (GLM)** that accommodates different error structures.

Initial Setup and Loading the Dataset in Stata

Before implementing any diagnostic tests, it is necessary to prepare the **Stata** workspace and load an appropriate dataset. For the purposes of this tutorial, we will utilize the **auto** dataset, which is a classic built-in **Stata** dataset containing various attributes of 1978 automobiles. This dataset is ideal for demonstrating normality tests because it contains variables with different distributional characteristics, such as **weight**, **mileage (mpg)**, and **engine displacement**. Accessing this data is straightforward and requires only a single command in the **Stata Command window**.

To begin, clear any existing data in the memory to avoid conflicts and then load the system dataset. This ensures that the environment is "clean" and that the variables we reference are exactly as expected. The following command retrieves the dataset from the **Stata** system directory and makes it active for analysis:

```
sysuse auto
```

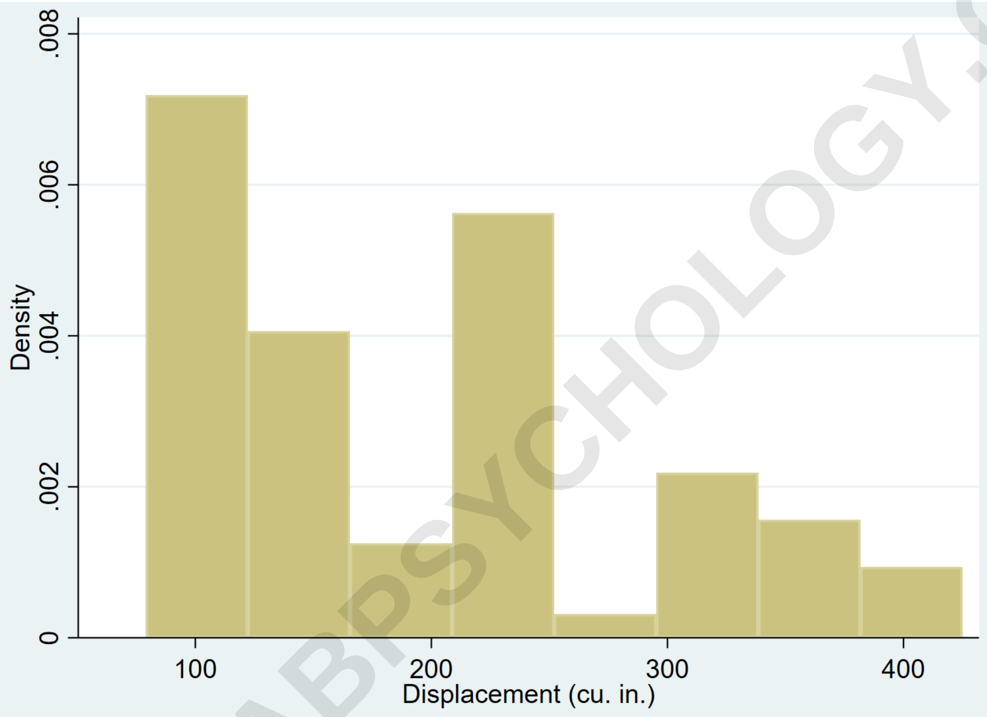
Once the dataset is loaded, it is good practice to use the `describe` and `summarize` commands to get a preliminary overview of the variables. The **displacement** variable, which measures engine size, is particularly interesting for normality testing because it often exhibits **skewness** in automotive data. By examining the **mean**, **standard deviation**, **minimum**, and **maximum** values, we can begin to form an intuition about the data's shape before applying more formal tests. This initial phase of **data cleaning** and exploration is a prerequisite for any meaningful **statistical analysis**.

Method 1: Visual Inspection Using Histograms

The most intuitive way to assess the distribution of a variable is through graphical visualization. A **histogram** partitions the data into discrete intervals (bins) and displays the frequency of observations within each bin. In a **normal distribution**, the **histogram** should be symmetric and bell-shaped, with the highest frequency of data points clustered around the **mean** and fewer observations appearing in the "tails" of the distribution. Any significant deviation, such as a long tail to the right or left, indicates a lack of normality.

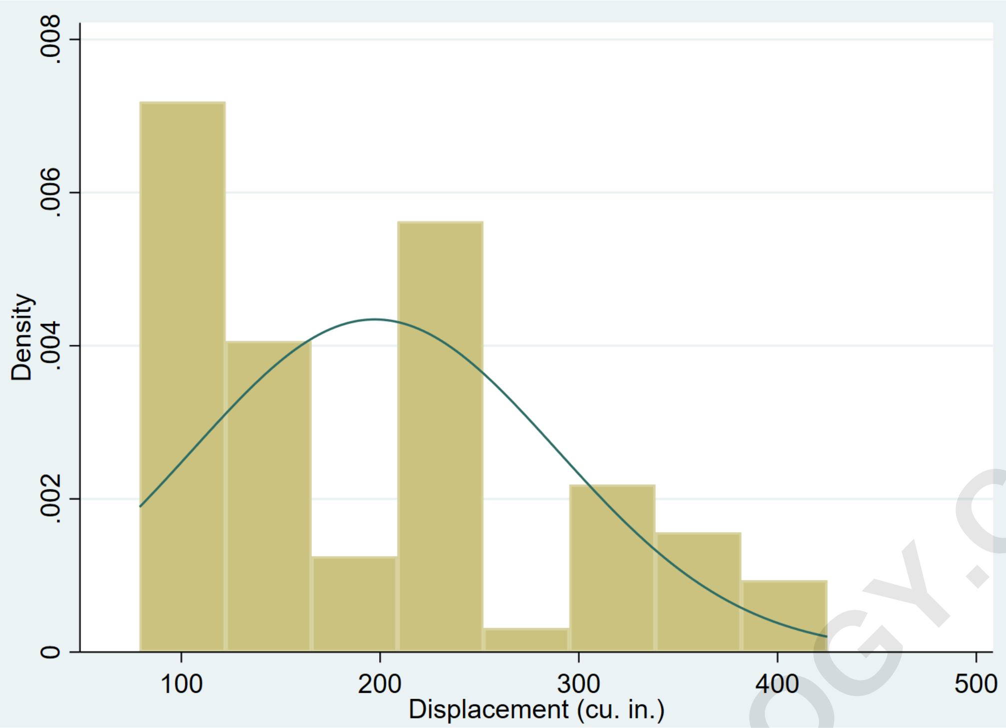
To generate a basic **histogram** for the **displacement** variable in **Stata**, we utilize the `hist` command. This command is highly customizable, allowing users to adjust bin widths or display percentages instead of raw frequencies. For our initial assessment, we run the following:

hist displacement



While a standard **histogram** provides a general sense of the data's shape, it can be difficult to judge normality without a reference line. **Stata** allows for the overlay of a theoretical **normal density curve** onto the histogram. This curve represents what the distribution would look like if it perfectly followed a Gaussian model with the same mean and standard deviation as your sample. Adding this curve makes it much easier to identify **skewness** or **kurtosis**. Execute the following command to see this comparison:

hist displacement, normal



Upon reviewing the output, it becomes evident that the **displacement** variable is not normally distributed. The data is **positively skewed** (right-skewed), as characterized by the cluster of values on the left side of the graph and a long tail extending toward the higher values on the right. While graphical methods are excellent for a quick "sanity check," they are subjective. Different bin sizes can change the appearance of the **histogram**, potentially leading to different interpretations. Consequently, formal **statistical hypothesis testing** is required to confirm these visual findings.

Method 2: The Shapiro-Wilk Test for Normality

The **Shapiro-Wilk test** is widely regarded as one of the most powerful and reliable formal tests for normality, particularly for small to moderate sample sizes. The **null hypothesis** (H0) for the test is that the sample data is drawn from a normally distributed population. If the **p-value** generated by the test is below a chosen **significance level** (commonly 0.05), we reject the **null hypothesis** and conclude that the data is not normally distributed. In **Stata**, this test is implemented using the `swilk` command, which is restricted to datasets containing between 4 and 2,000 observations.

To perform the **Shapiro-Wilk test** on the **displacement** variable, input the following command into the **Stata** console:

swilk displacement

```
. swilk displacement

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
displacement |         74    0.92542      4.803     3.423    0.00031
```

The output provides several key metrics that require careful interpretation. The **Obs** column shows that 74 observations were analyzed. The **W** statistic is the actual **Shapiro-Wilk** test statistic; values closer to 1.0 suggest normality. Most importantly, the **Prob > z** value represents the **p-value**. In this case, the **p-value** is approximately 0.00031. Since this value is significantly less than the 0.05 threshold, we have strong evidence to reject the assumption of normality for **displacement**.

One of the advantages of the `swilk` command is its ability to process multiple variables simultaneously. This is highly efficient when dealing with large datasets where several variables need to be checked before running a multivariate analysis. For example, to test **displacement**, **mpg**, and **length** in one go, use the following syntax:

swilk displacement mpg length

```
. swilk displacement mpg length

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
displacement |         74    0.92542      4.803     3.423    0.00031
         mpg |         74    0.94821      3.335     2.627    0.00430
      length |         74    0.97165      1.825     1.313    0.09461
```

In this multi-variable output, we observe that both **displacement** and **mpg** have **p-values** below 0.05, indicating non-normality. However, the **length** variable returns a **p-value** greater than 0.05. Therefore, for **length**, we fail to reject the **null hypothesis**, suggesting that its distribution does not significantly deviate from a normal distribution at the 5% **significance level**. This highlights how different variables within the same dataset can exhibit vastly different distributional properties.

Method 3: The Shapiro-Francia Test for Larger Samples

While the **Shapiro-Wilk test** is excellent, it has sample size limitations. The **Shapiro-Francia test** is an alternative that functions similarly but is designed to handle slightly larger datasets, specifically those with between 10 and 5,000 observations. It is essentially a modification of the **Shapiro-Wilk** procedure that simplifies the calculation of the test statistic, making it computationally efficient for larger **N**. Like its counterpart, the **Shapiro-Francia test** operates under the **null hypothesis** that the variable is normally distributed.

To execute this test in **Stata**, we use the `sfrancia` command. This is particularly useful when you have passed the 2,000-observation limit of `swilk` but still want a test based on the correlation between ordered observations and expected normal quantiles. To test the **displacement** variable, use:

sfrancia displacement

```
. sfrancia displacement

                   Shapiro-Francia W' test for normal data

    Variable |       Obs       W'          V'        z       Prob>z
-------------+------------------------------------------------------
displacement |        74    0.93011      4.975     3.110    0.00094
```

The interpretation of the **Shapiro-Francia** output mirrors that of the **Shapiro-Wilk**. In the results, the **W'** (W-prime) statistic serves as the test metric, and the **Prob > z** provides the **p-value**. For our **displacement** data, the **p-value** is 0.00094. Again, because this is well below 0.05, we conclude that the variable is not normally distributed. The consistency between the **Shapiro-Wilk** and **Shapiro-Francia** tests in this instance reinforces our conclusion regarding the data's non-normality.

It is important to note that as sample sizes become very large, almost any formal test will detect even trivial deviations from normality, leading to a rejected **null hypothesis**. This is a known phenomenon in **statistical hypothesis testing** where the test becomes "overpowered." In such cases, the **Shapiro-Francia test** remains a useful tool, but researchers should weigh the **statistical significance** against the practical significance by also consulting graphical plots. Efficiency and robustness make `sfrancia` a staple in the **Stata** user's toolkit for **econometric analysis**.

Method 4: The Skewness and Kurtosis Test (D'Agostino's K-squared)

Another powerful alternative in **Stata** is the **Skewness and Kurtosis test**, often accessed via the `sktest` command. This test is based on **D'Agostino's K-squared test** and evaluates normality by examining the **skewness** (asymmetry) and **kurtosis** (peakedness/heaviness of tails) of the data. A perfectly normal distribution has a **skewness** of 0 and a **kurtosis** of 3. The `sktest` calculates how much the sample's values deviate from these ideals and combines them into an aggregate **chi-squared test** statistic. This test requires a minimum of 8 observations to run effectively.

To run the **Skewness and Kurtosis test** for the **displacement** variable, enter the following:

sktest displacement

```
. sktest displacement

                   Skewness/Kurtosis tests for Normality
                                                        ——— joint ———
    Variable        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
─────────────┼──────────────────────────────────────────────────────────────
displacement          74      0.0337        0.2048        5.81        0.0547
```

The interpretation of `sktest` is slightly different as it provides individual **p-values** for **skewness** and **kurtosis**, followed by a combined probability. In our output, the **adj chi2(2)** is 5.81, and the **Prob > chi2** (the overall **p-value**) is 0.0547. Interestingly, at a strict 0.05 **significance level**, this test *fails* to reject the **null hypothesis**, as 0.0547 is slightly higher than 0.05. This result contrasts with the **Shapiro-Wilk** and **Shapiro-Francia** tests, illustrating how different statistical tests can yield varying results based on the specific characteristics of the data they emphasize.

The discrepancy observed here is a perfect example of why researchers should not rely on a single test. The `sktest` focuses on the third and fourth moments of the distribution, while the **Shapiro-Wilk test** is based on the correlation between the data and normal scores. When tests disagree, it usually suggests that the deviation from normality is subtle or specific to one aspect of the distribution (like the tails). Just like the other commands, `sktest` can be applied to multiple variables at once, providing a rapid way to check the **symmetry** and **tail-heaviness** of your entire dataset.

Choosing the Right Test and Interpreting Discrepancies

With multiple methods available in **Stata**, the question arises: which test should you trust? Generally, the **Shapiro-Wilk test** is preferred for its high **statistical power**, especially in smaller samples. However, the **Skewness and Kurtosis test** (`sktest`) provides valuable insight into *why* a distribution might be non-normal. For instance, if the **p-value** for **skewness** is low but the **kurtosis** **p-value** is high, you know the primary issue is asymmetry rather than extreme outliers. This distinction is crucial when deciding on a **data transformation** strategy.

Discrepancies between tests, as seen with our **displacement** variable where `swilk` rejected normality but `sktest` barely did not, often occur near the "borderline" of significance. In these situations, the visual evidence from **histograms** or **Normal Probability Plots (P-P plots)** should act as the tie-breaker. If the **histogram** shows a clear and significant lean, as it did for **displacement**, it is safer to treat the data as non-normal despite a borderline **p-value** from a specific test. **Statistical analysis** is as much an art as it is a science, requiring balanced judgment between formal output and visual patterns.

Furthermore, sample size plays a pivotal role in test selection. In very small samples (N < 20), normality tests often lack the power to detect even significant non-normality. Conversely, in very large samples, even a minuscule and irrelevant deviation from a **normal distribution** will result in a **statistically significant p-value**. Therefore, for large datasets, many statisticians prioritize graphical methods and the **Central Limit Theorem** over formal tests. Understanding these limitations ensures that the researcher does not overreact to **p-values** that may not have practical implications for the **robustness** of their final model.

Practical Steps for Handling Non-Normal Data

Once you have determined that your data is non-normal, the next step is deciding how to handle it. One common approach in **Stata** is **data transformation**. Applying a **logarithmic transformation** (`gen log_var = log(original_var)`) can often correct **positive skewness**, making the distribution more symmetric and closer to normal. Other options include the **square root transformation** for count data or the **inverse transformation** for extreme skewness. After transforming the data, it is essential to re-run the normality tests to verify that the transformation was successful.

If transformations fail to achieve normality, or if the research question requires the original scale of the variable, **non-parametric tests** may be the best alternative. These tests, such as the **Mann-Whitney U test** or the **Kruskal-Wallis test**, do not assume a specific underlying distribution and are based on the **ranks** of the data rather than the raw values. **Stata** has comprehensive support for these methods, which provide a valid way to conduct **hypothesis testing** without the stringent requirements of **parametric statistics**.

Finally, researchers may choose to use **robust standard errors** or **bootstrapping** techniques. **Bootstrapping** involves repeatedly resampling the data to estimate the **sampling distribution** of a statistic empirically, rather than relying on theoretical assumptions. This is particularly useful in **regression analysis** when residuals are non-normal. By using **Stata's** `vce(robust)` or `bootstrap` options, you can obtain valid **inference** even when the assumption of normality is violated. Ultimately, the goal is to ensure that the statistical conclusions are **reliable** and **replicable**, regardless of the initial shape of the data distribution.