Perform a Shapiro-Wilk Test in SAS

Authored by

Mohammed looti

November 1, 2025

Introduction: Assessing Data Distribution with the Shapiro-Wilk Test

The rigorous assessment of data distribution stands as a cornerstone of statistical analysis. Before applying many sophisticated parametric techniques, such as t-tests and ANOVA, analysts must first confirm whether their dataset conforms to a normal distribution. This crucial prerequisite ensures the validity of subsequent inferences. Among the various methods available for testing this assumption, the **Shapiro-Wilk test** is globally recognized for its exceptional power and reliability.

The Shapiro-Wilk test operates by comparing the observed cumulative distribution function of the sample data against the expected cumulative distribution function of a theoretical normal population. It is particularly well-suited for small to moderately sized samples (typically $n < 50$), although its application extends to larger datasets. A significant outcome from this test serves as a strong indication that the data deviates fundamentally from a Gaussian (normal) distribution.

This comprehensive tutorial offers a detailed, step-by-step methodology for correctly executing the Shapiro-Wilk test utilizing the industry-leading statistical software package, **SAS**. We will meticulously guide you through the process, covering essential data setup, the precise command syntax required, and, most importantly, the clear interpretation of the statistical output, providing a reliable framework for quantitative researchers.

The Critical Role of Normality Testing in Robust Statistical Inference

The assumption that underlying population data follows a normal distribution is foundational to many classical statistical inference procedures. If this assumption is ignored or violated, the resulting test statistics, confidence intervals, and overall statistical conclusions derived from parametric tests may be severely biased or entirely invalid. Consequently, failing to verify normality can lead researchers to erroneous conclusions based on misleading evidence.

Consider, for instance, the application of linear regression: verifying the normality of the residuals is essential for accurate model assessment. Similarly, in a Student's t-test, the data within each group must be normally distributed for the test results to hold true. When the normality assumption cannot be confirmed, researchers must pivot to non-parametric alternatives. While these alternatives are distribution-free, they often come at the cost of reduced statistical power, making the initial confirmation of normality via tests like the **Shapiro-Wilk test** indispensable.

Therefore, adhering to best practices in rigorous quantitative research necessitates employing formal statistical tests, such as the **Shapiro-Wilk test**, in conjunction with visual diagnostic tools like histograms and Q-Q plots. This preliminary analytical stage ensures that the chosen subsequent inferential approach is entirely appropriate and optimized for the specific characteristics of the data under examination.

Step 1: Efficient Data Preparation and Loading into SAS

The initial stage of any analysis in SAS involves defining and inputting the raw data into a structured dataset. For demonstration purposes, we will construct a small dataset containing 15 observations for a single continuous variable, which we name 'x'. This scenario mirrors a common research situation where a small sample must be analyzed to confirm its distributional characteristics before proceeding to more complex tests.

To achieve this, we employ the fundamental SAS DATA step, combined with the DATALINES statement, allowing us to input the scores directly into the statistical environment. This procedure effectively structures and populates the dataset, which we will formally name `my_data`, readying it for analysis.

The following code snippet clearly illustrates the required syntax for creating the dataset and then verifying its successful load using the `PROC PRINT` command. The clean, sequential structure of the code is emblematic of effective **SAS programming** practices, minimizing errors and maximizing replicability.

```
/*create dataset: 'my_data' contains 15 observations for variable 'x'*/
data my_data;
input x;
datalines;
3
3
4
6
7
8
8
9
12
14
15
15
17
20
21
;
run;

/*view dataset using PROC PRINT to ensure data integrity*/
```

```
proc print data=my_data;
```

The subsequent output (as shown in the image below) confirms that the dataset `my_data` has been accurately loaded into the SAS environment, displaying the 15 observations that will be subjected to the normality test.

Obs	x
1	3
2	3
3	4
4	6
5	7
6	8
7	8
8	9
9	12
10	14
11	15
12	15
13	17
14	20
15	21

Step 2: Executing the Shapiro-Wilk Test with PROC UNIVARIATE

With the data successfully loaded, the next phase involves executing the formal statistical procedure. In the **SAS** system, the procedure responsible for calculating descriptive statistics and tests related to distributional properties for a single variable is **PROC UNIVARIATE**. This procedure is incredibly powerful and provides a holistic view of the variable's characteristics.

To specifically instruct PROC UNIVARIATE to perform the Shapiro-Wilk test, we must include the mandatory `NORMAL` option within the procedure statement. This option triggers the calculation of a suite of normality tests, encompassing not only Shapiro-Wilk but also the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests, providing a comprehensive assessment of fit.

The syntax below is exceptionally concise, illustrating the minimal command structure necessary to generate the required normality statistics for the variable 'x' housed within the `my_data` file. This

efficiency is a hallmark of effective SAS analysis.

```
/*perform Shapiro-Wilk test using PROC UNIVARIATE and the NORMAL option*/
proc univariate data=my_data normal;
run;
```

Executing this straightforward command produces extensive output covering moments, quantiles, and extreme values. However, for our specific goal, we must narrow our focus exclusively to the table dedicated to formal tests of normality, demonstrating the skill required to efficiently extract critical information from detailed statistical reports.

Comprehensive Interpretation of the Normality Output

The output generated by PROC UNIVARIATE is rich in detail, but the section of paramount importance for our analysis is unequivocally the table labeled **Tests for Normality**. This section consolidates the results from the four key goodness-of-fit tests calculated by the procedure, each designed to assess the fit of the data to a normal distribution.

Within this pivotal table, analysts can locate the calculated test statistics and, critically, the associated p-values for every normality test. While all tests serve a similar purpose (evaluating the null hypothesis of normality), the **Shapiro-Wilk test** is typically prioritized, especially for smaller datasets, due to its documented superior statistical power.

The four primary tests for normality presented in the SAS output are:

The **Shapiro-Wilk Test**: Highly recommended for its power with small-to-medium sample sizes.
The Kolmogorov-Smirnov Test: Often applied to larger samples or when specific population parameters are not estimated from the sample.
The Cramer-von Mises Test: Offers an alternative measure of distributional fit, sometimes demonstrating higher sensitivity than Kolmogorov-Smirnov.
The Anderson-Darling Test: Provides increased weight to discrepancies observed in the tails of the distribution, making it sensitive to outliers.

By reviewing the output image below, we must locate the specific row corresponding to the **Shapiro-Wilk test** to obtain the W statistic and the corresponding critical p-value, which is the ultimate determinant of our conclusion regarding normality.

The UNIVARIATE Procedure
Variable: x

Moments			
N	15	Sum Weights	15
Mean	10.8	Sum Observations	162
Std Deviation	5.96657356	Variance	35.6
Skewness	0.30348727	Kurtosis	-1.1151814
Uncorrected SS	2248	Corrected SS	498.4
Coeff Variation	55.2460514	Std Error Mean	1.54056267

Basic Statistical Measures			
Location		Variability	
Mean	10.80000	Std Deviation	5.96657
Median	9.00000	Variance	35.60000
Mode	3.00000	Range	18.00000
		Interquartile Range	9.00000

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	7.010426	Pr > |t|	<.0001
Sign	M	7.5	Pr >= |M|	<.0001
Signed Rank	S	60	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.936921	Pr < W	0.3452
Kolmogorov-Smirnov	D	0.151886	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.052851	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.331213	Pr > A-Sq	>0.2500

Examination of this generated table reveals that the crucial p-value associated with the **Shapiro-Wilk test** is reported as **0.3452**. This single figure holds the key to determining whether our dataset meets the necessary assumption of normality.

Drawing Conclusions from the P-Value and Hypothesis Testing

Interpreting the result of the Shapiro-Wilk test requires a formal application of hypothesis testing

principles. The test is structured around two competing hypotheses:

H0 (The Null Hypothesis): The population data from which the sample was drawn follows a normal distribution.
HA (The Alternative Hypothesis): The population data does *not* follow a normal distribution.

The standard decision rule dictates that we compare the calculated p-value against a predetermined significance level ($alpha$), conventionally set at 0.05. If the p-value is less than the significance level ($p < 0.05$), we reject the null hypothesis, thereby concluding that the data significantly deviates from normality.

In the context of our specific example, the obtained p-value is **0.3452**. Since 0.3452 is considerably larger than the conventional $alpha$ of 0.05, we must consequently **fail to reject the null hypothesis**.

This result indicates that there is insufficient statistical evidence to conclude that our dataset is non-normal. Therefore, based on the rigorous findings of the **Shapiro-Wilk test**, researchers can confidently proceed with any parametric statistical methods--such as ANOVA or t-tests--that rely on the fundamental assumption of a normal distribution. Conversely, had the p-value been, for example, 0.005, the null hypothesis would have been rejected, necessitating the exploration of data transformations or the adoption of non-parametric analytical techniques.

Expanding Your SAS Statistical Toolkit

The successful performance and interpretation of the Shapiro-Wilk test in SAS represent a key step in mastering statistical analysis. Beyond normality testing, **SAS** provides a comprehensive suite of powerful tools for various inferential and descriptive procedures. Understanding the structure of these procedures allows researchers to build a highly efficient and versatile analytical workflow.

Analysts who frequently encounter distributional assumptions or need to conduct other common inferential tests will find the following procedural topics essential for expanding their analytical capability within the SAS environment:

Running T-tests (utilizing `PROC TTEST` for mean comparisons)
Conducting Analysis of Variance (employing `PROC GLM` or `PROC ANOVA` for comparing multiple groups)
Performing Correlation and Regression Analysis (using `PROC CORR` for associations or `PROC REG` for predictive modeling)

The following resources offer related tutorials explaining how to execute these other fundamental statistical procedures in SAS, complementing the knowledge gained here regarding the

assessment of the normal distribution assumption: