How to Perform a Kolmogorov-Smirnov Test in SAS to Check Data Distribution

Authored by

stats writer

December 1, 2025

The Kolmogorov-Smirnov Test, often abbreviated as the K-S test, is a powerful nonparametric statistical tool. In the context of data analysis using **SAS**, it is most frequently applied as a one-sample test to assess whether a given sample dataset is drawn from a specific theoretical probability distribution, such as the normal distribution. This assessment of sample normality is critical because many standard inferential statistical procedures (like t-tests and ANOVA) rely heavily on the assumption that the data follows a normal pattern. The core mechanism of the test involves comparing the observed Cumulative Distribution Function (CDF) of the sample data against the expected CDF of the theoretical distribution. The test statistic calculated is the maximum absolute difference between these two functions. A smaller maximum difference suggests a stronger agreement between the sample data and the hypothesized distribution, indicating that the data is likely drawn from that distribution. When checking for normality in **SAS**, the `PROC UNIVARIATE` command is the standard and recommended procedure for performing the Kolmogorov-Smirnov Test, although `PROC NPAR1WAY` can be used for the two-sample version.

The **Kolmogorov-Smirnov test** is fundamentally employed to determine whether a given data sample conforms to a specific theoretical distribution. Although it can compare two empirical samples, its most common use in descriptive statistics is assessing whether a sample is consistent with the assumptions of normality. This statistical inference allows researchers to decide which modeling techniques are appropriate for subsequent analysis, ensuring the validity of their conclusions.

This test is widely utilized because a great majority of classical statistical tests and procedures make the fundamental assumption that the underlying data distribution is normally distributed. Failing to confirm this assumption before proceeding with parametric tests can lead to inaccurate standard errors, skewed confidence intervals, and ultimately, incorrect interpretations of the study's results. Therefore, confirming the distributional properties of the data is a mandatory preliminary step in robust data analysis.

The following detailed, step-by-step example demonstrates precisely how to execute the Kolmogorov-Smirnov test on a single sample dataset within the **SAS** statistical software environment, focusing on the powerful capabilities offered by `PROC UNIVARIATE`.

Understanding the Kolmogorov-Smirnov Test Methodology

The Kolmogorov-Smirnov Test is defined by its reliance on the Cumulative Distribution Function (CDF). Unlike tests that rely solely on moments (like mean and variance), the K-S test considers the entire shape of the distribution. For a one-sample test, we calculate the empirical CDF, $F_n(x)$, which represents the proportion of observations in the sample that are less than or equal to x. This empirical function is then compared to the theoretical CDF, $F(x)$, which is defined by the hypothesized distribution (e.g., the standard normal distribution).

The critical element in the K-S test is the test statistic, D, which quantifies the largest vertical distance between the empirical CDF and the theoretical CDF across all possible values of x. Formally, $D = max_{x} |F_n(x) - F(x)|$. If the data truly follows the hypothesized distribution, this maximum deviation D should be small. A large D value suggests a significant discrepancy between the sample data and the reference distribution, leading to the rejection of the null hypothesis. The magnitude of D is thus a direct measure of the goodness-of-fit.

It is important to note that when testing for normality, the mean (mu) and standard deviation ($sigma$) of the reference normal distribution are typically estimated from the sample data itself. This procedure, known as using estimated parameters, slightly modifies the resulting critical values, making the test inherently conservative. While the K-S test is versatile, the Shapiro-Wilk test is often considered more powerful for detecting departures from normality, particularly for smaller sample sizes (n < 50). However, the K-S test remains a standard output in SAS's comprehensive univariate procedures.

Why Normality Testing is Crucial for Parametric Procedures

Statistical tests are broadly categorized into parametric and nonparametric methods. **Parametric tests**, which include staples like the t-test, Z-test, correlation analysis, and analysis of variance (ANOVA), assume specific distribution characteristics of the population from which the sample data is drawn. The most prevalent of these assumptions is normality. These tests derive their validity and power from the mathematical properties of the normal curve. When the data is indeed normally distributed, these tests provide the most efficient and reliable estimates and conclusions.

If the normality assumption is severely violated--meaning the data is heavily skewed or contains extreme outliers--the results of parametric tests can be highly misleading. Violations can inflate Type I error rates (false positives) or decrease the statistical power of the test, making it difficult to detect true effects. Furthermore, parameter estimates (like confidence intervals for the mean) rely on the assumption of symmetry inherent in the normal distribution. When this assumption fails, the interpretation of these parameters becomes compromised, potentially leading researchers to draw incorrect inferences about the population.

Therefore, the K-S test serves as a crucial diagnostic tool. If the Kolmogorov-Smirnov Test indicates a significant departure from normality (i.e., we reject the null hypothesis), the analyst must consider alternative strategies. These strategies often involve data transformations (such as logarithmic or square root transformations) to achieve a more normal distribution, or, more commonly, switching to **nonparametric equivalents**. Nonparametric tests, such as the Mann-Whitney U test or the Kruskal-Wallis test, do not require distributional assumptions and are robust to skewness and outliers, providing a safe alternative when normality cannot be established.

Applying the Kolmogorov-Smirnov Test in SAS

In the **SAS System**, the most straightforward and complete way to perform a one-sample Kolmogorov-Smirnov test for normality is by utilizing `PROC UNIVARIATE`. This procedure is designed to produce extensive descriptive statistics and detailed analyses of the distribution of a single variable, including several goodness-of-fit tests. While other procedures, such as `PROC NPAR1WAY`, can execute the two-sample K-S test, `PROC UNIVARIATE` provides the specific comparison against the theoretical normal distribution, which is the primary focus for checking test assumptions.

To implement the K-S test using `PROC UNIVARIATE`, the key step is the inclusion of the `HISTOGRAM` statement combined with the `NORMAL` option. The `NORMAL` option instructs SAS to overlay a normal density curve onto the histogram and, crucially, to perform a series of formal tests for normality, including the Kolmogorov-Smirnov Test, the Shapiro-Wilk test, and the Anderson-Darling test. By default, `PROC UNIVARIATE` estimates the population mean (mu) and standard deviation ($sigma$) from the sample data to define the parameters of the comparison normal distribution.

The precise syntax needed to invoke this functionality involves specifying `/ NORMAL(MU=EST SIGMA=EST)` within the `HISTOGRAM` statement. The `MU=EST` and `SIGMA=EST` parameters explicitly tell SAS to use the sample mean and standard deviation as the estimated parameters for the reference normal distribution. This ensures that the test correctly evaluates the goodness-of-fit against the closest possible normal curve that fits the sample data, maximizing the likelihood of accepting the null hypothesis if the data is close to normal.

Step-by-Step Example: Preparing the Sample Data in SAS

Before any statistical analysis can commence in **SAS**, the data must be properly defined and loaded into a dataset. For this example, we will simulate a modest dataset consisting of 20 observations ($n=20$). This sample size is small enough to demonstrate the mechanics clearly, yet large enough to provide meaningful output from the normality tests.

First, let's create a dataset named `my_data` in **SAS** using the `DATA` step and `DATALINES` statement, containing the sample size of $n = 20$ values:

```
/*create dataset*/
data my_data;
input Values;
datalines;
5.57
8.32
8.35
```

```
8.74
8.75
9.38
9.91
9.96
10.36
10.65
10.77
10.97
11.15
11.18
11.47
11.64
11.88
12.24
13.02
13.19
;
run;
```

This code block successfully defines a temporary SAS dataset named `my_data` and populates the variable `Values` with the 20 numerical entries. The structure adheres to standard **SAS programming practices**, employing the `DATA` and `INPUT` statements for definition, and concluding with the `RUN` statement to execute the step. The data values themselves appear to be centered around 10, suggesting a potential underlying normal distribution, but this assumption must be formally tested using the K-S procedure.

The creation of a clean dataset is the foundational prerequisite for all subsequent analysis. It ensures that `PROC UNIVARIATE` has the necessary structured input to calculate the empirical CDF and perform the necessary comparisons against the theoretical normal model. Once this step is executed, the dataset is loaded into memory, ready for statistical examination.

Executing the Normality Test using PROC UNIVARIATE

With the dataset prepared, the next step is to execute the procedure that performs the Kolmogorov-Smirnov Test. We will use `PROC UNIVARIATE`, specifying the dataset we just created, `my_data`. The procedure is powerful and generates numerous tables, but our primary focus here is the goodness-of-fit statistics output generated by the `HISTOGRAM` option.

Next, we'll use **proc univariate** combined with the appropriate options to perform a Kolmogorov-

Smirnov test to formally determine if the sample variable `Values` is normally distributed:

```
/*perform Kolmogorov-Smirnov test*/
proc univariate data=my_data;
histogram Values / normal(mu=est sigma=est);
run;
```

The syntax employed here is streamlined but effective. `PROC UNIVARIATE DATA=my_data;` invokes the procedure using our dataset. The key line is `histogram Values / normal(mu=est sigma=est);`. This command not only generates a visual histogram of the `Values` variable but, due to the `NORMAL` option, also generates a dedicated table titled "Tests for Normality." This table will contain the results for the Kolmogorov-Smirnov test, allowing us to evaluate the null hypothesis efficiently. Without the `NORMAL` option, **SAS** would only output descriptive statistics and general distribution plots, omitting the formal K-S test results required for this analysis. The explicit use of `MU=EST` and `SIGMA=EST` ensures that the comparison is against the normal distribution that best approximates our sample data.

Interpreting the Test Statistics and P-Value

After running the `PROC UNIVARIATE` code, **SAS** produces an extensive output file. Analysts must navigate this output to find the "Tests for Normality" table, where the specific results for the Kolmogorov-Smirnov Test are listed alongside other related tests. This table provides the calculated test statistic (D) and the corresponding p-value, which are the two critical components required for hypothesis testing.

At the bottom of the output section dedicated to normality tests, we can clearly observe the calculated test statistic and the corresponding p-value specifically for the Kolmogorov-Smirnov test, often labeled as D:

The UNIVARIATE Procedure
Fitted Normal Distribution for Values

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	10.375
Std Dev	Sigma	1.826721

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.10983186	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.04020411	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.29089867	Pr > A-Sq	>0.250

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	5.57000	6.12541
5.0	6.94500	7.37031
10.0	8.33500	8.03396
25.0	9.06500	9.14290
50.0	10.71000	10.37500
75.0	11.55500	11.60710
90.0	12.63000	12.71604
95.0	13.10500	13.37969
99.0	13.19000	14.62459

Based on the provided output snippet, the Kolmogorov-Smirnov test statistic, D, is calculated to be **0.1098**. Crucially, the corresponding p-value is reported as **>0.150**. This p-value is essential for determining the outcome of the hypothesis test. The D statistic represents the maximum vertical deviation between the empirical CDF of our sample data and the theoretical CDF of the fitted normal distribution. The p-value indicates the probability of observing a D statistic as extreme as 0.1098, assuming the null hypothesis (normality) is true. Since the p-value is bounded, but clearly large, we proceed to the final step of interpretation.

Formal Hypothesis Testing and Conclusion

To draw a valid conclusion from the K-S test results, we must formalize the testing framework by defining the null and alternative hypotheses. The Kolmogorov-Smirnov Test is designed to assess the fit of the sample distribution against a specified theoretical distribution, typically the normal distribution. Therefore, the hypotheses are stated as follows:

H0: The population from which the data sample was drawn is normally distributed. (The maximum difference D is small, suggesting a good fit.)

HA: The population from which the data sample was drawn is not normally distributed. (The maximum difference D is large, suggesting a poor fit.)

The critical decision rule in statistical testing relies on comparing the calculated p-value against a predetermined significance level ($alpha$), usually set at 0.05. If the p-value is less than $alpha$, we reject the null hypothesis, concluding that the data is significantly non-normal. Conversely, if the p-value is greater than $alpha$, we fail to reject the null hypothesis, concluding there is insufficient evidence to suggest the data deviates significantly from normality.

In our example, the calculated p-value from the **SAS** output is **>0.150**. Since this p-value (which is greater than 0.150) is substantially larger than the conventional significance level of $alpha = 0.05$, we **fail to reject the null hypothesis** (H_0). This outcome means that, based on the Kolmogorov-Smirnov Test, there is no statistically significant evidence at the 5% level to conclude that the sample data deviates from a normal distribution. Consequently, we can proceed with the assumption that the dataset is normally distributed, validating the use of parametric statistical methods for further analysis on the `Values` variable.

Further Resources and Related Tutorials

The ability to accurately assess the distributional assumptions of data is paramount for any rigorous statistical analysis. The use of `PROC UNIVARIATE` in **SAS** provides a comprehensive suite of tools, with the Kolmogorov-Smirnov Test being a cornerstone for normality checks. While SAS offers excellent facilities, understanding how to perform these fundamental tests across various software platforms is essential for modern data scientists.

For those utilizing different statistical software packages, the methodology and interpretation of the K-S test remain consistent, though the syntax will vary. Mastering the implementation of this and similar goodness-of-fit tests, such as the Anderson-Darling and Shapiro-Wilk tests, is crucial for maintaining the integrity and reliability of statistical models, regardless of the software environment.

The following tutorials explain how to perform a Kolmogorov-Smirnov Test in other statistical software: