Perform a Kolmogorov-Smirnov Test in SAS

Authored by

Mohammed looti

November 1, 2025

The **Kolmogorov-Smirnov test** (often abbreviated as the K-S test) is a crucial, non-parametric method used extensively in **statistical hypothesis testing**. Its primary application is to assess whether a given sample distribution significantly deviates from a specific theoretical distribution, most commonly the normal distribution. When applied to a single dataset, the K-S test provides a rigorous mechanism for determining if the underlying data is consistent with having been drawn from a **normally distributed** population.

The imperative need for this distributional check cannot be overstated in quantitative research. Many advanced parametric procedures--such as T-tests, Analysis of Variance (ANOVA), and Linear Regression--rely fundamentally on the assumption that the data or, more precisely, the residuals, follow a **normal distribution**. Failure to meet this core requirement, known as violating the assumption of **normality**, can severely compromise the validity of the analysis, leading to unreliable inferences and potentially flawed conclusions about the population parameters being studied.

This comprehensive tutorial provides a detailed, step-by-step guide on how to execute the **Kolmogorov-Smirnov test** using the robust statistical software package, **SAS**. We will navigate the process from initial data preparation and the specific procedure call using `PROC UNIVARIATE`, through to the critical interpretation of the diagnostic output generated by the **SAS** system.

The Mechanics of the Kolmogorov-Smirnov Test and Formal Hypotheses

The **Kolmogorov-Smirnov test** functions by systematically measuring the disparity between the empirical cumulative distribution function (ECDF) of the collected sample and the theoretical cumulative distribution function (CDF) of the reference distribution--in this tutorial, the standard normal distribution. The key result of this comparison is the test statistic, denoted as **D**, which is defined as the maximum absolute vertical difference observed between these two functions across all data points.

Interpreting the D statistic is straightforward: a smaller calculated D value signifies that the empirical distribution of the sample closely mirrors the theoretical normal distribution being tested. Conversely, a larger D statistic indicates a substantial and potentially significant departure from normality. While the K-S test serves as a reliable general-purpose goodness-of-fit test, it is important to note that for the specific task of evaluating normality, researchers often prefer or complement it with specialized alternatives, such as the Shapiro-Wilk test, particularly when working with smaller sample sizes where the K-S test might lack statistical power.

Prior to running any statistical analysis, it is mandatory to formally define the hypotheses that guide the decision-making process. The K-S test, when used for distributional assessment, operates under the following paired statements: the **null hypothesis** (H_0) and the alternative hypothesis (H_A).

H0: The underlying population from which the data was sampled is normally distributed.
HA: The underlying population distribution of the data is significantly different from the normal distribution.

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

To provide a clear and actionable demonstration of the K-S test, we must first establish a working dataset within the **SAS** environment. We will construct a simple dataset, conventionally named `my_data`, containing twenty observations ($N = 20$) for a single measured variable called `Values`. This instructional simulation allows us to precisely control the input data structure and subsequently observe how **SAS** processes the request for distributional analysis.

Data preparation in **SAS** is managed via the `DATA` step, which names the dataset, followed by the `INPUT` statement to define the variable names. The raw observations are then directly entered using the `DATALINES` statement. This streamlined structure ensures that our sample data is correctly formatted and ready for the rigorous statistical assessment of its distributional properties in the next analytical step.

The following **SAS** code block executes the data creation process. Notice the use of block comments (`/* */`)--a standard and essential practice for documenting and maintaining clarity in statistical programming projects, particularly when handling complex data transformations or multiple analytical steps.

```
/*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;

Executing the Normality Test Using PROC UNIVARIATE Syntax

With the dataset successfully loaded, the next critical step involves invoking the `PROC UNIVARIATE` procedure. This procedure is the cornerstone command in **SAS** for generating comprehensive descriptive statistics and conducting distributional assessments for continuous variables. To explicitly request the Kolmogorov-Smirnov test specifically for normality, we must incorporate the `HISTOGRAM` statement combined with the `NORMAL` option.

The precise syntax required is `histogram Values / normal(mu=est sigma=est)`. This command performs a multifaceted analysis. It not only generates a visual histogram of the `Values` variable but also overlays a best-fit theoretical normal density curve. Most importantly for our test, the parameters `MU=EST` and `SIGMA=EST` are essential; they instruct **SAS** to estimate the population mean (mu) and the population standard deviation ($sigma$) directly from our sample data. This methodology ensures that the K-S test evaluates the sample against the *most appropriate* normal distribution for that specific dataset.

Executing this single procedure call produces extensive statistical output, including the summary table that contains all the necessary test statistics required for formally testing the hypothesis against the assumption of normality. We will focus our attention on this specific output table in the next step to derive our conclusion.

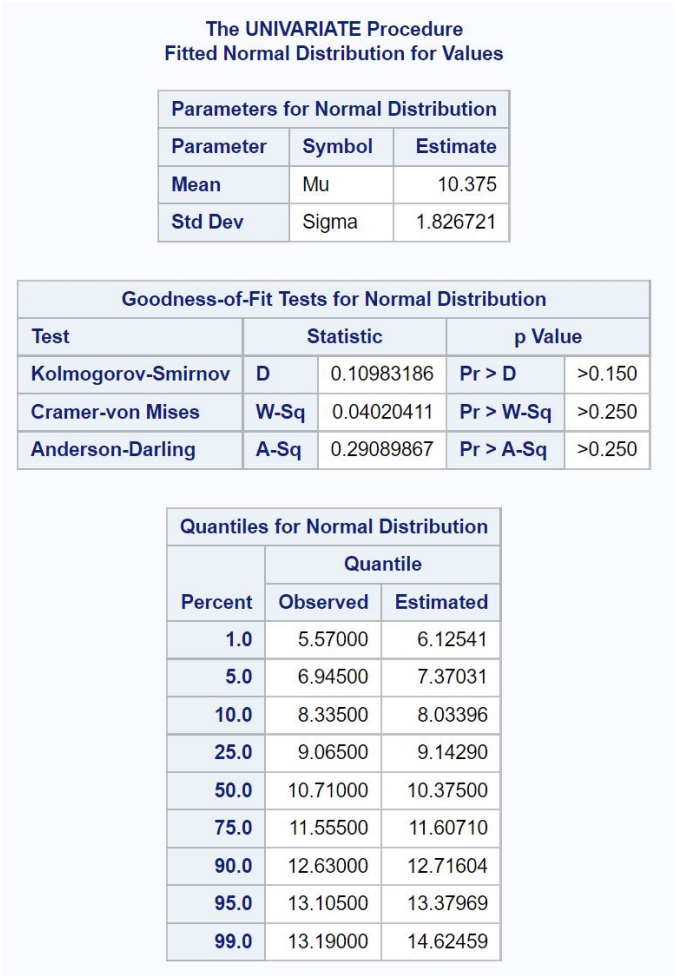
```
/*perform Kolmogorov-Smirnov test*/
proc univariate data=my_data;
histogram Values / normal(mu=est sigma=est);
run;
```

Interpreting the SAS Output, D Statistic, and P-Value

The results of the analysis are contained within the `PROC UNIVARIATE` output, specifically within a

dedicated section typically titled "Tests for Normality." This table presents a summary of the Kolmogorov-Smirnov test alongside other commonly used statistical tests for distributional adequacy. Our primary focus here is locating the calculated test statistic (D) and its corresponding critical p-value.

Careful review of the SAS generated output is necessary to extract these essential metrics. The visual representation below highlights the exact portion of the output where the goodness-of-fit test results are clearly displayed:



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

From the table, we ascertain two vital pieces of information: the K-S test statistic (D) is calculated as **0.1098**, and the corresponding asymptotic p-value is reported as **>0.150**. This p-value dictates our statistical decision concerning the null hypothesis (H_0). We traditionally compare this result against the standard significance level, $alpha$, which is set at 0.05 (or 5%) in most research domains.

The fundamental principle of frequentist hypothesis testing is applied here: if the calculated p-value is less than the chosen $alpha$ threshold, we must reject the null hypothesis. Since our obtained

p-value ($gt 0.150$) is substantially larger than the conventional $alpha = 0.05$, we conclude that there is insufficient statistical evidence to warrant the rejection of the null hypothesis.

Conclusion: Validating Assumptions for Parametric Statistical Tests

The decision to fail to reject the null hypothesis (H_0)--which posits that the data is normally distributed--is a positive outcome for the analyst. This outcome statistically confirms that the observed sample data does not exhibit significant deviations from the expected theoretical normal distribution. In practical terms, the data behaves in a manner consistent with having been sampled from a normally distributed population, validating a key distributional prerequisite.

This validation is an exceptionally crucial step in the analytical pipeline. Because the assumption of normality has been successfully confirmed by the K-S test, the analyst is now fully justified in proceeding with the application of robust parametric statistical tests that depend on this distribution, such as advanced regression modeling or t-tests. This prevents the necessity of resorting to complex data transformations or adopting less powerful non-parametric alternatives, thereby maximizing the statistical efficiency of the subsequent analysis.

By successfully executing and accurately interpreting the Kolmogorov-Smirnov test within the **SAS** environment, researchers ensure that the foundational assumptions of their chosen statistical models are met, leading directly to more robust, trustworthy, and reliable research conclusions.

Further Resources and Distributional Checks in Other Platforms

While this tutorial concentrated specifically on implementing the K-S test for normality using the **SAS** statistical software, the core principles governing distributional assessment remain universally applicable across all analytical platforms. Data scientists and analysts frequently use analogous commands and specialized procedures to conduct these essential checks in different software environments.

The following recommended resources offer supplementary guidance on how to perform the Kolmogorov-Smirnov test, the Shapiro-Wilk test, and similar critical distributional assessments in other popular statistical programming languages and software packages:

Detailed instructions on how to conduct Normality Checks in **R** using built-in functions such as `shapiro.test` and `ks.test`.
A guide to implementing Goodness-of-Fit Tests in **Python**, focusing on the powerful functionality provided by the SciPy library.
Tutorials demonstrating the use of **SPSS** for executing both the Kolmogorov-Smirnov and Shapiro-Wilk Tests within the comprehensive Explore command interface.