How do I Perform a Log Transformation in SAS?

Authored by

stats writer

November 19, 2025

A log transformation is a powerful mathematical technique used frequently in statistical analysis, particularly within SAS. This process involves replacing a variable 'X' with its natural logarithm, log(X). In SAS, this transformation is commonly executed either directly within a **DATA** step or using procedures like PROC UNIVARIATE, often specifying the **LOG** option. The primary goal of a log transformation is to handle skewed data, pulling the distribution closer to a **normal distribution**.

Achieving normality is critical because many standard statistical tests rely on this assumption to produce valid and accurate inferences. Besides PROC UNIVARIATE, advanced procedures like **PROC TRANSREG** can also be utilized for complex data transformations across various columns in a dataset.

Understanding the Need for Log Transformation

Most parametric statistical tests--such as t-tests, ANOVA, and linear regression--operate under the core assumption that the data for the variable of interest, or the residuals in a modeling context, follow a **normal distribution**. When data significantly violates this crucial assumption, the results derived from these tests can become unreliable, leading to inaccurate standard errors and potentially flawed conclusions about the population being studied.

Variables frequently encountered in real-world studies, such as income levels, certain biological concentrations, or population counts, often exhibit severe positive skewed data, where the majority of observations are concentrated on the low end and a long tail extends towards higher values. The log transformation serves as a powerful corrective measure for such non-normal, positively skewed distributions.

By compressing the large values disproportionately more than the smaller values, the logarithm function pulls the extreme right tail inwards. This non-linear operation effectively stabilizes the variance, aids in homogenizing residual error, and typically brings the distribution significantly closer to the bell-shaped curve required for valid parametric analysis. The decision to apply a log transform should always be grounded in empirical evidence derived from appropriate normality tests and careful visual inspection of the variable's distribution.

Initial Data Exploration and Testing for Normality in SAS

Before implementing any transformation, it is essential to establish baseline metrics for the variable's distribution. We begin by creating a simple dataset in SAS that is designed to demonstrate significant positive skewness. This dataset, which we name **my_data**, contains a single variable, **x**, characterized by a heavy concentration of low values and a few higher observations.

The initial steps involve defining the data using the **DATALINES** statement and then confirming its structure using **PROC PRINT**. This ensures that the data has been loaded correctly into the SAS environment before proceeding to the analytical phase. The code block below illustrates the data creation and immediate inspection:

```
/*create dataset*/
data my_data;
input x;
datalines;
1
1
1
2
2
2
2
2
2
3
3
3
6
7
8
;
run;

/*view dataset*/
proc print data=my_data;
```

The output confirms the structure of the variable **x**, showing 15 observations in total:

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

To rigorously assess the distribution, we utilize PROC UNIVARIATE. This procedure is the standard tool for generating descriptive statistics, producing graphical outputs like a histogram, and, most importantly, executing formal normality tests. We include the **NORMAL** option to ensure that SAS calculates the required test statistics.

The following code generates the necessary diagnostic output to determine if variable **x** is normally distributed:

```
/*create histogram and perform normality tests*/
proc univariate data=my_data normal;
histogram x;
run;
```

The execution of the procedure yields comprehensive output, including descriptive statistics and the results of various assessments:

The UNIVARIATE Procedure
Variable: x

Moments			
N	15	Sum Weights	15
Mean	3	Sum Observations	45
Std Deviation	2.20389266	Variance	4.85714286
Skewness	1.43206094	Kurtosis	0.96326857
Uncorrected SS	203	Corrected SS	68
Coeff Variation	73.4630887	Std Error Mean	0.56904264

Basic Statistical Measures			
Location		Variability	
Mean	3.000000	Std Deviation	2.20389
Median	2.000000	Variance	4.85714
Mode	2.000000	Range	7.00000
		Interquartile Range	1.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	5.272013	Pr > |t|	0.0001
Sign	M	7.5	Pr >= |M|	<.0001
Signed Rank	S	60	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.772253	Pr < W	0.0017
Kolmogorov-Smirnov	D	0.3	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.286181	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.512189	Pr > A-Sq	<0.0050

Evaluating the Original Distribution

The primary focus of this preliminary analysis lies in the table titled **Tests for Normality**. For this dataset, we specifically examine the results of the Shapiro-Wilk test, which is highly sensitive and appropriate for datasets with a small number of observations (N < 50). The test produces a p-value that is significantly less than the conventional alpha level of 0.05. This finding compels us to reject

the null hypothesis that the data is normally distributed, providing strong statistical evidence that the variable **x** is non-normal.

This statistical confirmation is visually corroborated by the histogram generated by PROC UNIVARIATE. The visualization clearly depicts a distribution that is heavily concentrated at the lower end of the scale, confirming the marked positive skewed data:

Given the combined weight of evidence from both the formal Shapiro-Wilk test and the visual representation in the histogram, we can confidently determine that the original data violates the normality assumption. Consequently, a log transformation is the appropriate next step to attempt to normalize the variable before proceeding with parametric statistical tests.

Step-by-Step Example: Implementing the Log Transformation

The most flexible and explicit method for performing a log transformation in SAS is by creating a new variable within a **DATA** step. This approach allows the user to maintain the original data while generating a transformed version for analysis. In this demonstration, we create a new dataset

named **log_data** based on the original **my_data**.

Within the **DATA** step, we utilize the built-in **LOG()** function. This function computes the natural logarithm (base e) for each observation of the variable **x**. For simplicity in this example, we overwrite the variable **x** in the new dataset with its log-transformed value, but best practice often involves creating a new variable name (e.g., **log_x**) to avoid confusion.

This transformation mathematically compresses the variance inherent in the original distribution. The code below executes this crucial step and uses **PROC PRINT** to inspect the newly calculated logarithmic values, demonstrating the structural change in the data:

```
/*use log transformation to create new dataset*/
data log_data;
set my_data;
x = log(x);
run;

/*view log transformed data*/
proc print data=log_data;
```

The resulting **log_data** dataset confirms that the original integer counts have been converted into decimal values representing the natural logarithm. Notice how the largest values (7 and 8) now have much smaller numerical differences between them than in the original scale, illustrating the variance-stabilizing effect of the log function:

Obs	x
1	0.00000
2	0.00000
3	0.00000
4	0.69315
5	0.69315
6	0.69315
7	0.69315
8	0.69315
9	0.69315
10	1.09861
11	1.09861
12	1.09861
13	1.79176
14	1.94591
15	2.07944

Analyzing the Transformed Data and Results

Following the successful application of the log transformation, it is absolutely essential to re-evaluate the distribution of the new variable. We must repeat the diagnostic steps performed earlier using the new dataset, **log_data**, to empirically confirm the effectiveness of the transformation. We once again invoke PROC UNIVARIATE with the **NORMAL** option and request a histogram for the transformed variable.

This comparative analysis will reveal how the statistical measures of normality have shifted. We are looking for the p-value of the normality tests to increase above the 0.05 threshold, indicating that the data is now statistically consistent with a normal distribution.

The code below executes the final diagnostic checks on the log-transformed data:

```
/*create histogram and perform normality tests*/
proc univariate data=log_data normal;
histogram x;
run;
```

The resulting output clearly shows a marked improvement in the statistical indicators of normality post-transformation:

The UNIVARIATE Procedure
Variable: x

Moments			
N	15	Sum Weights	15
Mean	0.88478874	Sum Observations	13.2718311
Std Deviation	0.65910349	Variance	0.4344174
Skewness	0.44759883	Kurtosis	-0.3719325
Uncorrected SS	17.8246104	Corrected SS	6.08184366
Coeff Variation	74.4927523	Std Error Mean	0.17017979

Basic Statistical Measures			
Location		Variability	
Mean	0.884789	Std Deviation	0.65910
Median	0.693147	Variance	0.43442
Mode	0.693147	Range	2.07944
		Interquartile Range	0.40547

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	5.199141	Pr > |t|	0.0001
Sign	M	6	Pr >= |M|	0.0005
Signed Rank	S	39	Pr >= |S|	0.0005

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.890891	Pr < W	0.0692
Kolmogorov-Smirnov	D	0.214383	Pr > D	0.0630
Cramer-von Mises	W-Sq	0.127487	Pr > W-Sq	0.0435
Anderson-Darling	A-Sq	0.718565	Pr > A-Sq	0.0480

Interpreting the Shapiro-Wilk test Results

The crucial evidence is found once again in the **Tests for Normality** table. For the log-transformed variable, the p-value associated with the Shapiro-Wilk test is now substantially greater than 0.05.

This result signifies that we no longer possess sufficient statistical evidence to reject the hypothesis of normality. Thus, the transformed variable is now suitable for use in parametric statistical tests.

Visually, the histogram confirms the statistical success. The distribution of the log-transformed data is considerably more symmetric and bell-shaped compared to the original skewed data. The excessive clustering at the low end has been dispersed, resulting in a distribution that more closely approximates the characteristics of a normal curve:

In summary, based on the quantitative evidence from the Shapiro-Wilk test and the qualitative assessment of the improved histogram, we conclude that the log transformation successfully created a variable that is much more normally distributed than the original variable, thereby validating its suitability for subsequent modeling.

Alternative Transformation Methods in SAS

While the **DATA** step provides the most direct means of creating a permanent log-transformed variable, SAS offers other powerful, context-specific procedures for data transformation. These alternatives are often employed in complex modeling scenarios where the transformation itself is part of a larger estimation process.

One such procedure is **PROC TRANSREG** (Transformation and Regression), which specializes in optimal scaling and non-linear transformation methods, including the family of Box-Cox transformations. **PROC TRANSREG** is exceptionally flexible, allowing users to simultaneously find the optimal transformation parameter that best linearizes relationships, stabilizes variance, or improves distributional properties for one or more variables within a regression or modeling framework.

Alternatively, for quick diagnostics without altering the dataset, PROC UNIVARIATE offers the **TRANSFORM=LOG** option. When this option is specified, the procedure calculates all descriptive statistics, tests for normality, and generates plots based on the temporary log-transformed values, providing a quick way to assess if the transformation would be beneficial before committing to a permanent change in the dataset.

The successful application of a data transformation is a fundamental skill in statistical computing. Mastering the use of the **LOG()** function within the **DATA** step ensures that researchers working in SAS can easily adjust skewed data to meet the stringent requirements of robust parametric analysis.