How to Normalize Data in SAS: A Step-by-Step Guide

Authored by

stats writer

December 1, 2025

Normalizing data in SAS is a fundamental and often mandatory preprocessing step in statistical analysis and machine learning. This process involves transforming the raw data to conform to a specific, standardized distribution. The most common form of this procedure is **Z-score standardization**, where the data is scaled such that the resultant set has an arithmetic mean of zero (0) and a standard deviation of one (1). This transformation is mathematically achieved by calculating the Z-score for every observation: subtracting the arithmetic mean from each original value and then dividing that result by the standard deviation of the entire dataset.

This rigorous standardization is crucial because it ensures that all variables contribute equally to any subsequent analytical procedure. Without standardization, features with larger inherent magnitudes or wider value ranges would disproportionately dominate distance calculations and model metrics, leading to biased results. By aligning the scales of different variables, standardization enhances comparability across the dataset, a property indispensable for advanced techniques such as cluster analysis, principal component analysis, and neural network training. This tutorial details the precise steps required to perform this critical data normalization efficiently using the robust `PROC STDIZE` procedure within the SAS statistical software suite.

In classical statistical methodology, to "normalize" a set of data values usually implies performing a scale transformation that results in a unit variance and a zero center. This scaling method is robust, highly interpretable, and widely utilized when preparing data for modeling, especially when the underlying distributions are roughly symmetrical. It is essential to understand the mathematical relationship between the raw scores and the resulting Z-scores to correctly interpret how extreme a standardized observation is relative to the rest of the sample population.

Understanding Data Normalization and Standardization

While general usage often treats the terms "normalization" and "standardization" synonymously, especially in introductory data science, the specific process implemented by SAS's Z-score method is formally known as **standardization**. Standardization centers the data around the mean (setting it to zero) and scales it using the standard deviation (setting it to one). This method differs from Min-Max normalization, which scales data strictly between 0 and 1, a method often sensitive to outliers. Standardization, by producing Z-scores, is generally preferred when working with variables that possess inherently different units, preventing arbitrary scaling differences from introducing bias into multivariate analyses.

The primary objective of standardization is to preserve the intrinsic shape of the original distribution while shifting and resizing its scale. If the original data follows a normal distribution, the standardized data will also be normally distributed, albeit as a standard normal distribution. This transformation converts every raw score into a metric--the Z-score--that immediately conveys how far, measured in units of standard deviation, an observation deviates from the central measure of

the dataset. This property makes the Z-score invaluable for outlier detection and hypothesis testing.

Why is Standardization Necessary for Analytical Models?

Data scaling, particularly Z-score standardization, resolves critical issues related to feature magnitude that plague many statistical and machine learning algorithms. When variables are measured on widely disparate scales--for example, comparing population size (in millions) with unemployment rates (as percentages)--algorithms that rely on calculating distances between data points, such as K-Nearest Neighbors, K-Means Clustering, or Principal Component Analysis, will inherently assign undue weight to the variable with the largest absolute range. This scale bias leads to inaccurate models and erroneous conclusions.

By applying normalization, we enforce scale invariance across all input features. Every feature, regardless of its original unit or range, now contributes equally to the calculation of distance and overall feature importance, assuming that all variables are judged to be equally relevant based on domain expertise. Furthermore, standardizing features to have a zero mean and unit variance significantly accelerates the convergence of optimization techniques used in machine learning, such as gradient descent. Centered inputs allow the optimization process to navigate the model's cost function space more directly and efficiently, avoiding issues like oscillation or excessively slow learning rates.

The Z-Score Standardization Formula Explained

The mathematical backbone of the standardization process in SAS is the Z-score formula, which transforms an original data point x into its standardized counterpart z. This calculation is non-linear and relies entirely on the derived statistics of the sample itself. The formula is universally defined as:

$$z = (x - bar{x}) / s$$

In this context, $bar{x}$ represents the sample mean of the entire dataset, and s denotes the sample standard deviation. It is crucial to use the sample statistics (as calculated from the training data) to transform both the training data and any future test or validation data, ensuring consistency across the entire modeling pipeline.

The resulting Z-score is an immediate indicator of a raw score's position relative to the center of the distribution, measured in standard deviation units. A Z-score of -1.5 indicates that the value is one and a half standard deviations below the mean, highlighting its relative position within the context of the dataset's overall spread. This metric provides unparalleled interpretability, making standardized data a powerful tool for researchers and analysts alike.

Setting up the Dataset in SAS: Step 1

To demonstrate this crucial procedure, we must first define and populate a sample dataset within the SAS environment. We create a simple dataset named `original_data` containing a single numeric variable, `values`, which will be the focus of our standardization. This initial data step is foundational and ensures the reproducibility of the entire normalization workflow.

We will utilize the `DATA` step combined with the `DATALINES` statement to input the sample values directly. Following the data input, we immediately employ the `PROC MEANS` procedure. This step is necessary to calculate and confirm the initial descriptive statistics--specifically the mean and standard deviation--which are the necessary parameters for the Z-score calculation.

We begin with the following sequence of raw observations for the variable `values`:

Data
12
14
15
15
16
17
18
20
24
25
26
29
32
34
37

The following code executes the data creation and displays the unstandardized metrics:

```
/*create dataset*/
data original_data;
input values;
datalines;
12
14
```

```
15
15
16
17
18
20
24
25
26
29
32
34
37
;
run;

/*view mean and standard deviation of dataset*/
proc means data=original_data Mean StdDev ndec=3;
var values;
run;
```

Reviewing Initial Statistics

After running the initial SAS code block, the output from `PROC MEANS` confirms the descriptive properties of the raw data. This statistical summary is essential because it reveals the original scale and spread of the data before any transformation is applied.

The summary table generated by `PROC MEANS` shows the following statistics:

The MEANS Procedure

Analysis Variable : values	
Mean	Std Dev
22.267	7.968

From this output, we clearly establish that the raw dataset possesses an arithmetic mean ($\bar{x}$) of **22.267** and a standard deviation (s) of **7.968**. These two parameters are the exact values that `PROC STDIZE` will use to calculate the Z-scores. Our primary objective in the subsequent step is to transform the data such that a renewed calculation of these two statistics

yields 0 and 1, respectively.

Executing Standardization with PROC STDIZE (Step 2)

The most robust and streamlined approach to implementing Z-score standardization in SAS is leveraging the `PROC STDIZE` procedure. This procedure is specifically designed for scaling and centering data, offering a range of standardization methods. By default, and most commonly, it performs the Z-score transformation.

We use the `PROC STDIZE` statement, specifying the input data using the `DATA=` option and defining a new output dataset using `OUT=normalized_data`. The `VAR` statement identifies the variables (in this case, `values`) that must be transformed. Following the standardization step, we include `PROC PRINT` to visually inspect the new standardized scores, and then re-run `PROC MEANS` on the `normalized_data` dataset. This final statistical check confirms that the transformation has successfully yielded a mean of 0 and a standard deviation of 1, validating the entire process.

```
/*normalize the dataset*/
proc stdize data=original_data out=normalized_data;
var values;
run;

/*print normalized dataset*/
proc print data=normalized_data;

/*view mean and standard deviation of normalized dataset*/
proc means data=normalized_data Mean StdDev ndec=2;
var values;
run;
```

Interpreting the Standardized Data (Step 3)

The final output demonstrates the successful standardization of the dataset. The table contains the original index, the original raw values, and the newly calculated standardized values (Z-scores).

Obs	values
1	-1.28842
2	-1.03743
3	-0.91194
4	-0.91194
5	-0.78644
6	-0.66094
7	-0.53545
8	-0.28446
9	0.21753
10	0.34302
11	0.46852
12	0.84501
13	1.22149
14	1.47248
15	1.84897

The MEANS Procedure

Analysis Variable : values	
Mean	Std Dev
0.00	1.00

The crucial confirmation comes from the final `PROC MEANS` output: the mean of the normalized dataset is verified to be **0**, and the standard deviation is confirmed to be **1**. This signifies a successful implementation of the Z-score standardization. The resulting values in the `values` column of `normalized_data` are the Z-scores, which provide immediate context regarding each observation's position.

To reinforce the interpretation, let us recall the transformation formula:

Normalized value $z = (x - bar{x}) / s$

Consider the data point $x = 12$ from our original dataset, where $bar{x} = 22.267$ and $s = 7.968$. The corresponding standardized value is:

Normalized value $z = (12 - 22.267) / 7.968 approx **-1.288**$

This Z-score of -1.288 tells us that the original value of 12 is situated **1.288 standard deviations below the mean** of the dataset. Conversely, any positive standardized score, such as 1.849 (corresponding to the raw value 37), indicates that the observation is 1.849 standard deviations above the mean. Standardized values close to zero represent observations near the average, while large absolute values indicate potential outliers or extreme observations relative to the sample average.

Further Exploration of SAS Utilities

While this demonstration focused on Z-score standardization (the default method for `PROC STDIZE`), the procedure is highly versatile and capable of executing various other normalization and scaling techniques. For instance, if the goal is to scale data into a restricted range (e.g., between 0 and 1), the `METHOD=RANGE` option can be used. Furthermore, for datasets known to contain significant outliers, analysts might opt for more robust methods, such as centering based on the median and scaling based on the interquartile range (IQR), to mitigate the undue influence of extreme values.

The core principle remains constant: data preparation through standardization is a non-negotiable step for achieving reliable and unbiased results in complex statistical modeling tasks. Mastering the use of `PROC STDIZE` is crucial for maintaining data integrity and maximizing the predictive power of analytical models developed in SAS.

The following tutorials explain how to perform other common statistical tasks in SAS: