How to Calculate the Median Using PROC MEANS in SAS

Authored by

stats writer

November 19, 2025

The PROC MEANS procedure in SAS is an indispensable and robust tool widely utilized for calculating various descriptive statistics across large and complex datasets. Analysts frequently rely on this procedure to gain immediate insights into the central tendency and dispersion of their variables. One of the most critical measures of central tendency is the median. Although powerful, understanding the nuances of PROC MEANS, particularly how to explicitly request the median, is essential for comprehensive data summarization.

The median is defined as the middle value in a dataset when that dataset is sorted in ascending or descending order. It effectively divides the data distribution into two equal halves. Unlike the arithmetic mean, the median is highly resistant to the influence of outliers or extreme values, making it a preferred measure of central tendency for skewed distributions. Including the median in your output provides a much clearer picture of the typical value, especially when the mean might be misleading due to non-normal data shapes.

This detailed guide will serve as an expert resource for SAS users, demonstrating precisely how to modify the standard PROC MEANS syntax to ensure the median value is included in the resulting summary table. We will explore the default statistics, contrast them with the enhanced output, and provide practical code examples illustrating the technique in a real-world scenario involving athletic performance data. Mastering this simple syntax adjustment unlocks deeper and more accurate statistical reporting capabilities within the SAS environment.

Introduction to PROC MEANS and the Need for the Median

The core utility of the PROC MEANS procedure is to generate concise and readable summary reports. It is one of the most frequently executed procedures in any SAS workflow, offering quick calculation of fundamental measures for specified variables. When initiating statistical analysis, the first step often involves running this procedure to understand the basic characteristics of the variables, such as sample size, minimum, maximum, and average values. These preliminary statistics help in quality checking the data and preparing for more advanced modeling.

However, many new SAS users are surprised to find that the default execution of PROC MEANS excludes the median. While the mean provides an average, relying solely on it can be misleading if the data distribution is heavily skewed--a common occurrence in fields like finance, epidemiology, or quality control. Therefore, the ability to explicitly request and display the median alongside the mean becomes paramount for achieving robust and responsible statistical reporting. The median offers a measure of central location that is less susceptible to distortion caused by unusually high or low observations, thereby providing a more faithful representation of the data's true center.

To successfully integrate the median into the summary statistics output, the user must explicitly list the desired statistical keywords within the procedure call. This is achieved by specifying options on

the main PROC MEANS statement. While this may seem like a minor technical detail, understanding this mechanism is key to customizing the output for specific analytical needs. We will elaborate on this syntax requirement in detail, ensuring that users can consistently and reliably generate outputs that include this crucial measure of central tendency.

Understanding Default Behavior in PROC MEANS

By default, when you execute the PROC MEANS procedure without specifying any statistical keywords, it automatically calculates a standard set of descriptive statistics. This standard set is engineered to provide a rapid overview of the data characteristics. These defaults typically include the total count of non-missing observations (**N**), the arithmetic average (**Mean**), the dispersion measure (**Standard Deviation** or Std Dev), and the boundaries of the data (**Minimum** and **Maximum**). This core set is sufficient for many basic data exploration tasks but deliberately omits less common or computationally intensive statistics like the median or mode.

The statistics generated by default are computationally efficient and widely applicable. They establish the groundwork for subsequent analysis. For instance, knowing the minimum and maximum helps detect data entry errors, while the standard deviation indicates the typical distance observations lie from the mean. However, the exclusion of the median means that if the underlying distribution is highly skewed, the reported mean might present a misleading picture of the "center" of the data. Analysts must be aware of this default limitation to avoid drawing incorrect conclusions based solely on the standard output table.

To confirm this behavior, consider running the basic syntax: simply calling **PROC MEANS** and specifying a variable without any options. The resulting output table will display the standard five statistics. To obtain a complete set of summary measures, especially if robust statistics are required, the user must explicitly instruct PROC MEANS to calculate the median using the appropriate keyword. This customization capability highlights the flexibility and power of the SAS environment, allowing users to tailor their reports precisely to their analytical needs.

The Significance of the Median in Data Analysis

The median is often referred to as the 50th percentile, representing the value below which 50% of the observations fall. Its primary advantage over the mean lies in its robustness against extreme values. For example, when analyzing household income, a small number of extremely wealthy individuals (outliers) can significantly inflate the mean income, making it appear higher than what the typical household actually earns. In contrast, the median income remains unaffected by these extremes, providing a truer measure of central location for the general population. This characteristic makes the median indispensable in fields dealing with natural, financial, or social data, where skewness is frequently observed.

Furthermore, comparing the median and the mean offers immediate insight into the distribution shape of a variable. If the mean and median are approximately equal, the distribution is likely symmetric (e.g., normally distributed). If the mean is significantly greater than the median, the data is positively (right) skewed, indicating a few high outliers. Conversely, if the mean is less than the median, the data is negatively (left) skewed. This simple comparison is a fundamental step in exploratory data analysis (EDA) and helps determine whether subsequent parametric tests, which often assume normality, are appropriate.

For data scientists and statisticians working in SAS, generating both the mean and the median simultaneously using PROC MEANS ensures that the summary statistics provide a complete and balanced view of the central tendency. By making the effort to include the median, analysts enhance the integrity and defensibility of their statistical summaries, moving beyond basic arithmetic averages to measures that reflect the true pattern of the underlying data distribution.

Implementing the MEDIAN Option: Essential Syntax

To instruct PROC MEANS to calculate and display the median, we must explicitly use the keyword **MEDIAN** in the PROC statement. Unlike procedures where options are supplied after a slash (*/*), the statistical keywords in PROC MEANS are listed immediately following the dataset specification, before the semicolon. If you wish to retain the default statistics while adding the median, you must list all desired keywords explicitly, as SAS overrides the default output list entirely once any option is specified.

The standard syntax structure for achieving this comprehensive statistical output is as follows. We specify the procedure name, the dataset, and then list the required statistical keywords. A common list includes the total number of observations (**N**), the arithmetic average (**Mean**), the median (**Median**), the measure of spread (**Std** for Standard Deviation), and the range boundaries (**Min** and **Max**). This combination provides a holistic view of the variable.

The following syntax block illustrates the correct implementation of the **MEDIAN** option, combining it with other frequently requested statistics. This robust command ensures that the output table contains all necessary components for initial data assessment, allowing for rapid comparison between the mean and the median.

```
proc means data=my_data N Mean Median Std Min Max;
var points;
run;
```

This particular example calculates the total number of observations, mean, median, standard deviation, minimum, and maximum value for a variable specifically named **points**. This detailed

specification ensures that the resulting report is both comprehensive and easy to interpret, containing all measures vital for a preliminary descriptive statistics overview.

Practical Demonstration: Setting up the Sample Dataset

To fully illustrate the utility of the **MEDIAN** option within PROC MEANS, we will establish a sample dataset in SAS. This dataset simulates information collected about various basketball players, tracking their performance metrics such as points scored and assists made. Creating a controlled dataset allows us to verify the calculated median manually and observe the effect of the procedure modifications directly. We utilize the **DATA** step and **DATALINES** statement for swift data entry.

The structure of the dataset includes three variables: **team** (character variable representing the team), **points** (numeric variable tracking points scored by the player), and **assists** (numeric variable tracking assists). The data intentionally contains a variety of values, some of which might act as minor outliers, thereby making the comparison between the mean and the median particularly relevant for the **points** variable.

We begin by executing the following code block to create and then display our sample dataset, named **my_data**. The **PROC PRINT** step confirms the successful loading and structure of the data, which serves as the foundation for our subsequent statistical calculations.

```
/*create dataset*/
data my_data;
input team $ points assists;
datalines;
A 10 2
A 17 5
A 17 6
A 18 3
A 15 0
B 10 2
B 14 5
B 13 4
B 29 0
B 25 2
C 12 1
C 30 1
C 34 3
C 12 4
C 11 7
```

```
;
run;

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

The visualization below confirms the structure of the dataset, showing 15 observations across the three variables, ready for analysis using the PROC MEANS procedure.

Example: Display Median in PROC MEANS in SAS

Obs	team	points	assists
1	A	10	2
2	A	17	5
3	A	17	6
4	A	18	3
5	A	15	0
6	B	10	2
7	B	14	5
8	B	13	4
9	B	29	0
10	B	25	2
11	C	12	1
12	C	30	1
13	C	34	3
14	C	12	4
15	C	11	7

Executing PROC MEANS Without and With the MEDIAN Option

Before demonstrating the solution, let us first confirm the default behavior of PROC MEANS. We will calculate the summary statistics for the **points** variable using the minimal syntax, expecting to see the standard five statistics and noting the absence of the median. This exercise establishes a clear baseline against which the enhanced output can be compared, highlighting the precise impact of adding the statistical keyword.

The code below executes the standard PROC MEANS procedure on our created dataset, focusing exclusively on the **points** variable. Note that no statistical options are specified on the PROC statement itself:

```
/*calculate summary statistics for points variable*/
proc means data=my_data;
var points;
run;
```

Upon reviewing the output generated by the default execution, as displayed in the image below, it is evident that PROC MEANS provides the following descriptive statistics:

N: The total number of observations
Mean: The mean value of points
Std Dev: The standard deviation of points
Minimum: The minimum value of points
Maximum: The maximum value of points

Crucially, notice that the median value is not included in this default output, confirming the necessity of modifying the syntax to obtain this specific measure.

The MEANS Procedure

Analysis Variable : points				
N	Mean	Std Dev	Minimum	Maximum
15	17.8000000	7.8848861	10.0000000	34.0000000

Integrating the Median: The Customized Solution

Now, we implement the required solution by adding the **MEDIAN** keyword to the PROC MEANS statement. Since we want a complete picture, we also list the keywords for N, Mean, Std, Min, and Max. This specific listing overrides the procedure's default behavior and forces the computation and display of all specified statistics, including the median.

The following syntax block executes the enhanced procedure. We explicitly request the desired statistics, ensuring a comprehensive summary report for the **points** variable:

```
/*calculate summary statistics for points and include median value*/
proc means data=my_data N Mean Median Std Min Max;
```

```
var points;
run;
```

Upon reviewing the resulting output, as depicted below, the summary table is significantly enhanced. The output now includes a column labeled **Median**, providing the middle value of the distribution alongside the other critical measures. This modification successfully addresses the requirement of displaying the median in the PROC MEANS output.

The MEANS Procedure

Analysis Variable : points					
N	Mean	Median	Std Dev	Minimum	Maximum
15	17.8000000	15.0000000	7.8848861	10.0000000	34.0000000

Interpreting the Enhanced Summary Statistics

With the customized output now generated, we can analyze the complete set of descriptive statistics for the **points** variable. A crucial observation is the calculated median value, which turns out to be **15**. Comparing this to the **Mean**, which is approximately **17.8**, we immediately notice a difference. Since the mean is greater than the median, this suggests that the distribution of points scored is positively (right) skewed. This skewness is likely driven by the few high-scoring players (with 29, 30, and 34 points), which pull the mean upwards while having no effect on the position of the middle observation (the median).

The standard deviation, which is roughly 7.72, indicates a relatively high degree of dispersion in the scores. However, the median (15) provides the analyst with a more accurate representation of the typical performance level for the players in this sample. If an executive or manager were only given the mean (17.8), they might overestimate the general performance level. Providing both the mean and the median offers a much more nuanced and statistically responsible summary.

In conclusion, incorporating the **MEDIAN** option into the PROC MEANS procedure is straightforward yet profoundly impactful. It transforms a standard statistical summary into a comprehensive descriptive statistics report, enabling analysts using SAS to accurately assess central tendency, particularly when dealing with non-normal data distributions. Understanding and utilizing this simple syntax enhancement is a hallmark of efficient and expert SAS programming.

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