Learning Descriptive Statistics with SAS: A Comprehensive Guide

Authored by

Mohammed looti

November 16, 2025

The Foundational Role of Descriptive Statistics in SAS

Descriptive statistics form the fundamental core of rigorous data analysis, providing immediate, actionable numerical summaries that efficiently characterize the essential features of any given **dataset**. These critical metrics reveal the data's underlying structure, addressing key aspects such as central tendency (where data points converge), variability (the extent of data dispersion), and the overall shape of the distribution. Crucially, descriptive analysis focuses exclusively on the observed sample, summarizing its characteristics without attempting to make broader inferences about a population. This ensures an unbiased and clear initial assessment before progressing to more complex inferential methodologies.

In professional statistical computing environments, particularly within **SAS**, calculating these initial descriptive summaries is not optional; it is a mandatory preliminary step. This exploration allows expert analysts to quickly validate data quality, pinpoint potential **outliers**, and establish the necessary foundational context required for selecting appropriate statistical models or formulating hypotheses. A deep understanding of a variable's basic statistics--such as its **mean** and **standard deviation**--is indispensable for accurately interpreting the results of any advanced analysis that follows.

SAS provides highly optimized, specialized procedures designed to compute these metrics with maximum efficiency. This comprehensive guide will detail the two core procedures utilized by analysts to generate descriptive statistics for numeric variables, highlighting their specific applications and the distinct output they produce:

The highly efficient **PROC MEANS**, which is tailored for rapid calculation of core summary statistics including count (N), mean, and standard deviation.

The exhaustive **PROC UNIVARIATE**, which delivers a much deeper statistical profile, encompassing detailed moments, precise percentiles, and formal tests for distribution normality.

Preparing Your Sample Data for SAS Analysis

Before initiating any statistical procedure in **SAS**, the critical first step involves defining and cleaning the input **dataset**. To effectively illustrate the capabilities of **PROC MEANS** and **PROC UNIVARIATE**, we will start by constructing a simple, sample dataset named **my_data**. This dataset is structured to capture performance metrics, containing the categorical grouping variable **team**, along with the two primary numeric variables: **points** and **assists**. This straightforward configuration allows us to easily demonstrate both overall summary calculations and detailed group-wise analyses.

The following **SAS** code block initializes and populates this sample data structure. The process

begins with the **DATA step**, which formally establishes the structure of **my_data** and defines the variable types--for example, **team** is designated as a character variable using the $ sign, while **points** and **assists** are defined as numeric. Subsequently, the **DATALINES** statement is used for the direct entry of fifteen raw data points. To ensure the successful creation and accuracy of the input data, a basic **PROC PRINT** procedure is immediately executed, displaying the full contents of the new dataset in the output log.

```
/*create dataset: my_data*/
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 structure and values*/
proc print data=my_data;
```

The visual confirmation presented below verifies the structure of our newly constructed sample dataset. It explicitly validates the presence of the three defined variables (team, points, assists) and confirms the fifteen individual **observations** that will be used as input for calculating the subsequent descriptive statistical analyses detailed throughout this tutorial.

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

Calculating Essential Summary Statistics with PROC MEANS

The **PROC MEANS** procedure is an indispensable utility for core data analysis within the **SAS** ecosystem. It provides the most efficient methodology for generating basic yet crucial summary statistics for numeric variables. Its primary strength lies in its speed and ease of use, delivering essential metrics such as the count (N), the **mean**, the **standard deviation**, the **minimum**, and the **maximum** values. Analysts rely on **PROC MEANS** as the default choice when a rapid, high-level snapshot of a variable's central location and overall spread is required for initial data assessment.

To illustrate its simplicity, we will calculate these core summary metrics exclusively for the **points** variable within our **my_data** **dataset**. The necessary syntax is remarkably minimal: the procedure is called, the input dataset is specified, and the variable(s) of interest are explicitly designated using the **VAR statement**. It is worth noting that if the **VAR statement** were omitted, the procedure would default to analyzing all numeric variables available within the specified dataset, providing an aggregate summary for each.

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

The execution of this concise code generates an immediate, standardized output table, shown below. This table presents the essential **descriptive statistics** required for an initial grasp of the **points** variable's characteristics, confirming the effective sample size (N) and clearly defining the central point and the boundaries of the data distribution.

The MEANS Procedure

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

The standard output columns generated by **PROC MEANS** include critical measures such as: **N** (the count of non-missing **observations**); the **Mean**, which serves as the measure of central tendency; the **Std Dev** (**standard deviation**), which quantifies the data's dispersion relative to the mean; the **Minimum** value, indicating the lowest recorded score; and the **Maximum** value, representing the highest score observed in the data.

Performing Comparative Group Analysis using the CLASS Statement

In practical data analysis scenarios, it is often necessary to move beyond simple aggregate summaries and investigate how descriptive metrics vary across distinct subgroups. The **CLASS statement** is the primary mechanism within **PROC MEANS** for facilitating this crucial comparative group analysis. By specifying a categorical variable (or classification variable) within the **CLASS statement**, the procedure is instructed to calculate separate, independent sets of summary statistics for the designated numeric variables based on every unique level found in that categorical factor.

To illustrate this powerful segmentation capability, we will analyze how the summary statistics for the **points** variable differ among the various teams (A, B, and C) recorded in our **my_data dataset**. The following code block utilizes the **CLASS statement** to designate **team** as the grouping variable. This analytical approach provides a granular view, enabling analysts to directly compare performance metrics between groups rather than relying solely on a single, potentially misleading, overall average.

```
/*calculate summary statistics for points, grouped by team*/
proc means data=my_data;
class team;
var points;
```

run;

The resulting output, clearly visible in the image below, elegantly organizes the results into distinct rows corresponding to each unique team level. This structured presentation facilitates immediate comparative analysis, showing the **mean**, **standard deviation**, **minimum**, and **maximum** points scored individually for Team A, Team B, and Team C. Such detailed, group-specific insights are crucial for performance assessment, helping to rapidly identify which group exhibits the highest average performance or the greatest degree of score variability.

The MEANS Procedure

Analysis Variable : points						
team	N Obs	N	Mean	Std Dev	Minimum	Maximum
A	5	5	15.4000000	3.2093613	10.0000000	18.0000000
B	5	5	18.2000000	8.2885463	10.0000000	29.0000000
C	5	5	19.8000000	11.2338773	11.0000000	34.0000000

In-Depth Data Exploration with PROC UNIVARIATE

While **PROC MEANS** is excellent for providing rapid, essential statistical summaries, the **PROC UNIVARIATE** procedure is the definitive tool when a much deeper, exhaustive statistical profile of a numeric variable is required. **PROC UNIVARIATE** significantly extends the analysis beyond basic measures, generating extensive output that meticulously documents detailed percentiles, moments (crucial measures of distribution shape), and formal tests for adherence to specific distributions. It is an indispensable utility for thorough data investigation, quality assessment, and exploratory data analysis, particularly when understanding the exact shape, skewness, and **kurtosis** of a variable's distribution is paramount.

To showcase its comprehensive capabilities, we apply **PROC UNIVARIATE** to the **points** variable from our **my_data** dataset. The command structure is deliberately similar to **PROC MEANS**, requiring only the dataset specification and the variable of interest, which is defined using the mandatory **VAR statement**. This syntactical simplicity belies the extraordinary richness and depth of the statistical detail generated upon execution.

```
/*calculate detailed descriptive statistics for points variable*/
proc univariate data=my_data;
var points;
run;
```

The resulting output is extensive, typically encompassing multiple tables that meticulously chart every aspect of the data distribution. The images provided below offer a representative view of this generated output for the **points** variable, underscoring the vast array of statistical measures calculated in a single procedure run. This includes all standard measures of location and variability, alongside advanced diagnostics related to distribution shape and moments.

The UNIVARIATE Procedure
Variable: points

Moments			
N	15	Sum Weights	15
Mean	17.8	Sum Observations	267
Std Deviation	7.88488608	Variance	62.1714286
Skewness	1.00931793	Kurtosis	-0.2991564
Uncorrected SS	5623	Corrected SS	870.4
Coeff Variation	44.2971128	Std Error Mean	2.03586883

Basic Statistical Measures			
Location		Variability	
Mean	17.80000	Std Deviation	7.88489
Median	15.00000	Variance	62.17143
Mode	10.00000	Range	24.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	8.743196	Pr > |t|	<.0001
Sign	M	7.5	Pr >= |M|	<.0001
Signed Rank	S	60	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	34
99%	34
95%	34
90%	30
75% Q3	25
50% Median	15
25% Q1	12
10%	10
5%	10
1%	10
0% Min	10

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10	6	18	4
10	1	25	10
11	15	29	9
12	14	30	12
12	11	34	13

The comprehensive statistics provided by **PROC UNIVARIATE** cover three principal analytical categories: central tendency (such as the **mean**, **median**, and **mode**); dispersion (including **standard deviation**, **variance**, **range**, and **interquartile range**); and distributional shape (skewness and kurtosis). Furthermore, it meticulously details various related values, including the **minimum** and **maximum** scores, thereby delivering a holistic statistical summary essential for deep data exploration.

Granular Distribution Analysis: Grouping Data with PROC UNIVARIATE

For analytical scenarios that require the detailed statistical depth of **PROC UNIVARIATE** alongside subgroup analysis, the powerful **CLASS statement** is fully integrated and supported. Utilizing the **CLASS statement** allows analysts to generate a complete, multi-page univariate profile--including all moments, percentiles, and normality tests--for a numeric variable, broken down independently by every unique level of a specified categorical variable. This feature is absolutely crucial when seeking to understand subtle differences in distribution characteristics and data quality between

distinct groups.

We can extend our previous analysis by instructing **PROC UNIVARIATE** to apply its exhaustive calculations to the **points** variable, segmented by the **team** variable. The code below demonstrates this necessary grouping, utilizing the **VAR statement** for specifying the analysis variable and the **CLASS statement** for defining the grouping factor. This operation generates a highly detailed, comprehensive comparative report.

```
/*calculate detailed descriptive statistics for points, grouped by team*/
proc univariate data=my_data;
class team;
var points;
run;
```

Upon execution, this code yields extensive output that is systematically segmented into distinct, self-contained reports for Team A, Team B, and Team C. Each section provides a full set of univariate **descriptive statistics**, allowing for a precise comparison of central tendency (e.g., **mean** and **median**) and variability (e.g., **variance** and **interquartile range**) specific to each team. This granular level of analysis is invaluable for identifying group-specific performance patterns, detecting heterogeneity, and guiding targeted analytical interventions or subsequent inferential modeling.

Conclusion: Selecting the Right SAS Procedure

The ability to accurately calculate and correctly interpret **descriptive statistics** forms the essential, non-negotiable foundation of effective data analysis. Within the **SAS** environment, analysts are equipped with two powerful and distinct procedures, **PROC MEANS** and **PROC UNIVARIATE**, each optimized for different depths and stages of data exploration.

PROC MEANS is the tool of choice when efficiency and conciseness are prioritized. It excels in providing a rapid overview of key summary figures, such as the mean, standard deviation, minimum, and maximum values. It is exceptionally efficient for preliminary data profiling and quick comparisons, especially when enhanced by the use of the **CLASS statement** for generating group-wise summaries.

Conversely, **PROC UNIVARIATE** is specifically engineered for analytical depth. It generates an expansive statistical portfolio that includes the **median**, **mode**, **variance**, detailed percentiles, and crucial distribution diagnostics like skewness and kurtosis. This procedure should be employed whenever a comprehensive, in-depth understanding of a single variable's distribution shape, moments, and overall data quality is essential prior to engaging in inferential modeling.

By mastering the appropriate application and correct interpretation of these two distinct procedures, analysts gain the capability to effectively summarize and characterize any **dataset**. This disciplined approach establishes a rigorous and reliable groundwork for all subsequent statistical analysis, hypothesis testing, and data-driven decision-making processes. We strongly encourage readers to continue practicing and exploring the diverse range of optional statements available within both procedures to maximize their utility across various analytical contexts.

Additional Resources for SAS Proficiency

To continue building and refining your proficiency in the **SAS** environment, the following tutorials provide guidance on performing other common and advanced data processing tasks: