

How to Easily Calculate Deciles in SAS Using the NTILE Function

Authored by
stats writer

November 22, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Easily Calculate Deciles in SAS Using the NTILE Function*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=99707>

Calculating deciles is a fundamental task in quantitative analysis, allowing researchers to divide a distribution into ten equally sized groups. While SAS offers several powerful tools for this purpose, understanding the nuances of functions like the NTILE function and procedures like PROC UNIVARIATE is essential for generating accurate and meaningful results. The initial approach sometimes considered involves the NTILE function, which is primarily designed to assign observations to specified groups based on their rank.

The NTILE function accepts two core parameters: the input **data set** and the desired number of groups, or tiles. For instance, if you apply NTILE to a data set containing 10 observations and specify 2 as the tile count, the function effectively performs a median split, classifying the first five values into the first tile and the remaining five into the second. However, for calculating the precise decile boundary values themselves, the preferred and statistically rigorous method in SAS utilizes specialized procedures, which we will explore in depth.

In the field of statistics, **deciles** are formally defined as the nine points that partition a ranked data set into ten segments, with each segment containing exactly 10% of the total frequency of observations. This categorization method is invaluable for summarizing large distributions, identifying central tendencies, and detecting outliers or skewness within the data structure.

Specifically, deciles represent particular points along the distribution:

The first decile (D1) marks the point below which **10%** of all observed data values reside.

The second decile (D2) indicates the threshold where **20%** of all data values fall below it, and this progressive pattern continues through D9, which accounts for 90% of the data. Accurate calculation of these percentile points is crucial for robust reporting and comparative analysis.

Understanding Deciles and Quantiles in Data Analysis

Quantiles are markers used in statistics to divide the range of a probability distribution into continuous intervals with equal probabilities, or to divide the observations in a sample in the same way. Deciles are merely a specific type of quantile. While quartiles divide data into four parts and percentiles divide data into 100 parts, deciles offer a balanced, 10-part division, making them highly effective for large-scale data segmentation and distributional summary.

Analyzing decile values allows analysts to rapidly assess the spread of data and understand where specific observations fall relative to the overall population. For instance, knowing that a certain score falls in the 8th decile immediately tells us that 80% of the population scored lower, providing context that a raw score alone cannot offer. This makes decile analysis particularly useful in applications such as finance (analyzing income distribution) and educational assessment (benchmarking test scores).

When working with statistical software like SAS, the methodology chosen for calculating these quantiles must be precise. Using grouping functions (like NTILE) often provides rank categorization rather than the exact boundary values (the decile points) themselves. For formal statistical reporting, calculating the precise boundary values is necessary, requiring specialized procedures that handle interpolation and boundary conditions accurately.

The Role of PROC UNIVARIATE in Decile Calculation

In SAS, the most reliable and widely accepted method for calculating the exact decile values--the points that divide the distribution--is by using the **PROC UNIVARIATE** procedure. This procedure is designed for comprehensive analysis of a single variable, including descriptive statistics, distribution fitting, and, critically, quantile calculation. By leveraging specific statements within PROC UNIVARIATE, we can instruct SAS to compute the nine specific percentile values corresponding to the decile boundaries (D1 through D9).

PROC UNIVARIATE calculates quantiles based on a range of definition methods, ensuring that the resulting decile values are mathematically sound and align with standard statistical practices. This precision is paramount when accuracy is required. Unlike simpler grouping methods, PROC UNIVARIATE provides the actual numerical value that separates one decile group from the next, offering actionable thresholds for decision-making.

To successfully implement this procedure for decile calculation, users must clearly define three components within the PROC step: the input data set, the variable of interest, and the specific percentile points to be outputted. This structured approach guarantees a transparent and replicable statistical analysis process, which is essential for auditability and validation in professional data science projects.

Essential SAS Syntax for Decile Calculation

You can use the following standard syntax pattern to calculate the deciles for any continuous variable within a **data set** in SAS. This uses the OUTPUT statement within PROC UNIVARIATE to create a new data set containing only the calculated quantile values.

/*Calculate decile values for variable called var1 using PROC UNIVARIATE*/

proc univariate data=original_data;

var var1;

output out=decile_data;

pctlpts = 10 to 100 by 10

pctlpre = D_;

run;

This block of code outlines the necessary steps to transition from raw data to derived decile statistics. It first calls the specialized **PROC UNIVARIATE** procedure, specifies the variable to be analyzed (`var1`), and then uses the `OUTPUT` statement to direct the results into a new **data set** named `decile_data`. The subsequent statements are dedicated entirely to defining the desired quantiles.

Note: The `pctlpts` statement specifies exactly which percentiles should be calculated. Since deciles correspond to the 10th, 20th, 30th, up to the 90th percentile (D9), specifying `10 to 100 by 10` requests all these points, including the maximum (100th percentile). The `pctlpre` statement specifies the prefix that SAS will apply to the new variables created in the output **data set** (in this case, `D_10`, `D_20`, `D_30`, etc.). This ensures clarity when interpreting the resulting table.

Deconstructing the PCTLPT and PCTLPRE Statements

The `PCTLPT` statement, short for "percentile points," is the instruction that tells SAS exactly which cumulative percentages along the distribution we are interested in. For standard deciles, we aim for the 10%, 20%, 30%, and so on. The highly efficient syntax `10 to 100 by 10` automatically generates this sequence, saving the analyst from manually listing all ten values. Including 100 ensures the calculation includes the maximum value of the variable, often treated as the 10th decile boundary.

It is important to understand that the values returned by the `PCTLPT` statement are the actual scores or measurements from the input variable. For example, if the `PCTLPT` is 30, the output value (`D_30`) is the score at which 30% of the observations fall below it. This method relies on interpolation if the exact percentile point does not align perfectly with an observed data point, thus providing a true quantile value rather than just the closest observed value.

The `PCTLPRE` statement, or "percentile prefix," is crucial for data housekeeping and readability. When SAS executes the quantile calculation, it needs to name the new variables in the output data set. By setting `pctlpre = D_`, we enforce a consistent naming convention where the resulting variables are instantly recognizable as decile results (e.g., `D_10`, `D_50`, `D_90`). Without a defined prefix, SAS might use less descriptive system-generated names, complicating subsequent reporting and data merging operations.

Practical Example: Setting Up the Dataset

To demonstrate the robust capabilities of `PROC UNIVARIATE`, we will use a small sample **data set** containing player performance data, measuring the points scored across two hypothetical teams. The variable of interest for our decile calculation will be `points`.

The following code defines this sample data set using the standard `DATA` step and `DATALINES`

statement, followed by a procedural step to visualize the raw data structure and confirm successful loading into the SAS environment:

```
/*Create dataset named original_data*/
```

```
data original_data;
```

```
input team $ points;
```

```
datalines;
```

```
A 12
```

```
A 15
```

```
A 16
```

```
A 21
```

```
A 22
```

```
A 25
```

```
A 29
```

```
A 31
```

```
B 16
```

```
B 22
```

```
B 25
```

```
B 29
```

```
B 30
```

```
B 31
```

```
B 33
```

```
B 38
```

```
;
```

```
run;
```

```
/*View the raw dataset structure using PROC PRINT*/
```

```
proc print data=original_data;
```

```
run;
```

The resulting data set contains 16 total observations (player scores). This collection of scores represents the variable distribution upon which the decile calculation will be performed. SAS will internally sort the points variable values (12, 15, 16, 16, 21, 22, 22, 25, 25, 29, 29, 30, 31, 31, 33, 38) to determine the precise cut-off values for the ten segments, adhering to the principles of distributional statistics.

Obs	team	points
1	A	12
2	A	15
3	A	16
4	A	21
5	A	22
6	A	25
7	A	29
8	A	31
9	B	16
10	B	22
11	B	25
12	B	29
13	B	30
14	B	31
15	B	33
16	B	38

Executing the Decile Calculation Code

With the `original_data` set successfully created and verified, the next logical step is to execute the core statistical procedure. The following code utilizes PROC UNIVARIATE to calculate the decile boundaries for the `points` variable and then displays the results using PROC PRINT.

```
/*Calculate decile values for the points variable*/
```

```
proc univariate data=original_data;
```

```
var points;
```

```
output out=decile_data
```

```
pctlpts = 10 to 100 by 10
```

```
pctlpre = D_;
```

```
run;
```

```
/*View the resulting deciles stored in decile_data*/
```

```
proc print data=decile_data;
```

```
run;
```

Upon execution, the PROC UNIVARIATE procedure processes the `points` data. It determines the nine boundary values that partition the scores into ten equal frequency groups, effectively

calculating the 10th, 20th, 30th, up to the 100th percentile. These results are then meticulously stored in the `decile_data` output data set, where the column names are prefixed with 'D_' as instructed by the `PCTLPRE` statement.

The subsequent `PROC PRINT` command is absolutely necessary to visualize the generated data set, transforming the calculated numbers into an easily interpretable table. This output represents the final outcome of the decile calculation process, providing the analyst with the key threshold values needed for distributional analysis.

Interpreting the Output and Decile Values

The output table generated displays the computed decile values for the `points` variable across a single row, as quantiles are static measures of the entire distribution. Each column header (`D_10`, `D_20`, etc.) corresponds to a specific percentile calculated, where the associated numerical value is the score threshold.

Obs	D_10	D_20	D_30	D_40	D_50	D_60	D_70	D_80	D_90	D_100
1	15	16	21	22	25	29	30	31	33	38

Here is a detailed interpretation of the key decile values presented in the output table, which allows for immediate insights into the player performance distribution:

The value of the first decile (D1, or 10th percentile), designated by `D_10`, is **15**. This signifies that 10% of the scores in the data set are less than or equal to 15 points.

The value of the second decile (D2, or 20th percentile), labeled `D_20`, is **16**. This indicates that 20% of the players scored 16 points or less.

The value of the third decile (D3, or 30th percentile), displayed as `D_30`, is **21**. We can conclude that 30% of the observations fall below the score of 21.

The value of the fourth decile (D4, or 40th percentile), shown as `D_40`, is **22**. This means that 40% of the player scores are 22 or below.

The progression continues: the fifth decile (D5, 50th percentile, or the **median**) provides the exact midpoint score, and so on. The ninth decile (D9, 90th percentile, `D_90`) defines the threshold for the top 10% of scores, which is crucial for identifying top performers or potential outliers. Understanding these thresholds is essential for effective performance benchmarking.

Comparing NTILE and PROC UNIVARIATE for Data Segmentation

As noted earlier, while this methodology focuses on using `PROC UNIVARIATE` to calculate the

boundary values, it is often confused with the NTILE function. The primary purpose of the NTILE function, typically employed within a `DATA` step, is to assign an ordinal group index (1 through 10 for deciles) to each observation based on its rank relative to others, partitioning the data set into ten bins of approximately equal size.

The fundamental distinction is output type: `PROC UNIVARIATE` yields the specific numerical score that acts as the cut-off point (e.g., the score of 15 points is the D1 boundary), whereas the NTILE function creates a new categorical variable containing an integer (1 to 10) indicating which decile group the observation belongs to. For precise statistical measurement and reporting of percentile values, `PROC UNIVARIATE` is the appropriate and reliable tool.

In contrast, if the analyst's goal is purely data segmentation--for example, ranking employees into performance brackets for bonus allocation--the NTILE function is often the more direct and computationally efficient method for generating those group assignments directly onto the original data set. The choice depends entirely on whether the analyst requires the **boundary values** or the **group assignment**.

Summary of Best Practices

To ensure robust and accurate decile calculation in SAS, analysts should adhere to several key best practices. Always use **PROC UNIVARIATE** with the `PCTLPT` and `PCTLPRE` statements when the objective is to determine the precise numerical thresholds that divide the data distribution. Avoid relying solely on `PROC MEANS` or `PROC RANK` for these specific quantile points, as they may lack the precision offered by UNIVARIATE's dedicated quantile algorithms.

Furthermore, careful naming convention through the `PCTLPRE` statement is critical for maintainability and collaboration. A clear prefix like `D_` ensures that generated results are immediately understandable by anyone reviewing the code or the output data set. Finally, always include a verification step, such as `PROC PRINT`, immediately after the calculation to visually inspect the output and confirm that the decile values align logically with the source data distribution.

Mastering this technique allows analysts to move beyond basic descriptive statistics and engage in sophisticated distributional analysis, providing deeper insights into data concentration, variance, and segment performance across various quantitative measures.