CHAPTER

41

The UNIVARIATE Procedure

Overview

The UNIVARIATE procedure provides data summarization tools, high-resolution graphics displays, and information on the distribution of numeric variables. For example, PROC UNIVARIATE

- calculates descriptive statistics based on moments
- calculates the median, mode, range, and quantiles
- calculates the robust estimates of location and scale
- calculates confidence limits

- tabulates extreme observations and extreme values
- generates frequency tables
- plots the data distribution
- performs tests for location and normality
- performs goodness-of-fit tests for fitted parametric and nonparametric distributions.
- creates histograms and optionally superimposes density curves for fitted continuous distributions (beta, exponential, gamma, lognormal, and Weibull) and for kernel density estimates
- creates quantile-quantile plots and probability plots for various theoretical distributions and optionally superimposes a reference line that corresponds to the specified or estimated location and scale parameters for the theoretical distribution
- creates one-way and two-way comparative histograms, comparative quantile-quantile plots, and comparative probability plots
- insets tables of statistics in the graphical displays (high-resolution graphs)
- creates output data sets with requested statistics, histogram intervals, and parameters of the fitted distributions.

Output 41.1 on page 1319 shows a default univariate analysis for student exam scores. The statements that produce the output follow:

```
options pagesize=36;
proc univariate data=score;
run;
```

By default, the tests for location examine the hypothesis that the mean is equal to zero. Optionally, you can request a test for the hypothesis that the mean is equal to a specified value μ_0.

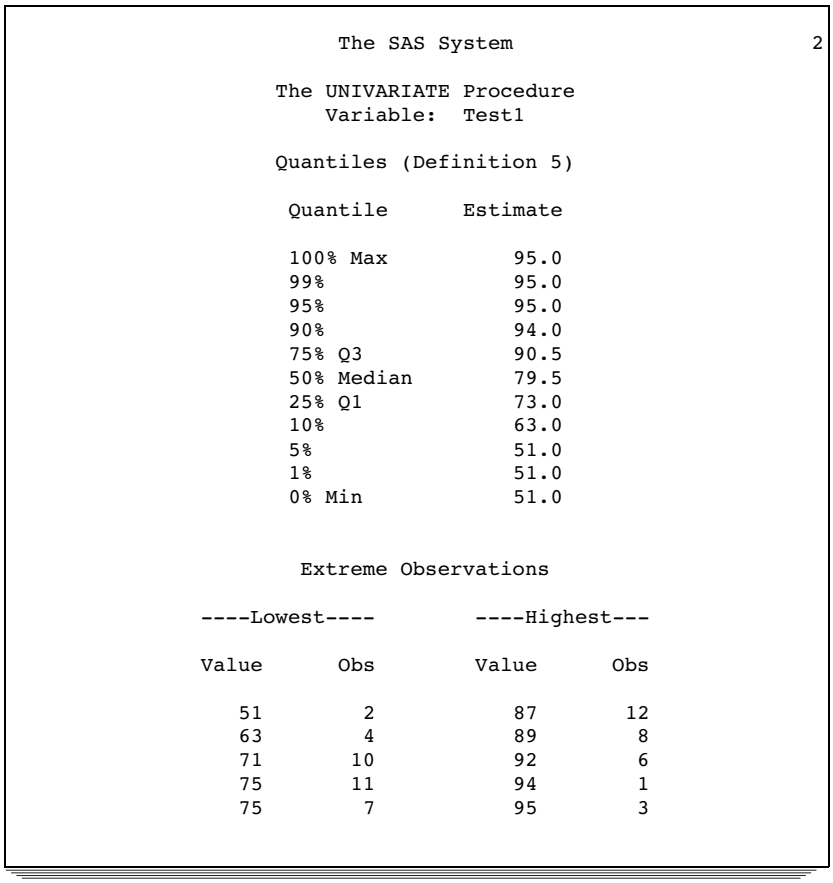
Output 41.2 on page 1321 and Output 41.3 on page 1324 are the result of a more extensive univariate analysis. The analysis examines the data distribution of student exam scores and creates an output data set that saves percentiles that were not computed by default. The statements that produce the analysis also

- specify the null hypothesis for the tests for locations
- perform tests for normality
- plot the data distribution
- specify the analysis variables
- request confidence limits for parameters and quantiles
- list the five highest and lowest extreme values
- print an output data set that contains percentiles.

For an explanation of the program that produces both these reports, see Example 5 on page 1432.

Output 41.1 The Default Univariate Analysis

```
                          The SAS System                          1

                    The UNIVARIATE Procedure
                        Variable:  Test1

                             Moments

N                          12    Sum Weights                 12
Mean                    79.25    Sum Observations           951
Std Deviation      13.3152339    Variance            177.295455
Skewness           -0.7841891    Kurtosis            0.27709746
Uncorrected SS          77317    Corrected SS           1950.25
Coeff Variation     16.801557    Std Error Mean      3.84377695


                  Basic Statistical Measures

         Location                      Variability

     Mean     79.25000     Std Deviation           13.31523
     Median   79.50000     Variance               177.29545
     Mode     75.00000     Range                   44.00000
                           Interquartile Range     17.50000


                  Tests for Location: Mu0=0

        Test           -Statistic-    -----p Value------

        Student's t    t  20.61774    Pr > |t|    <.0001
        Sign           M         6    Pr >= |M|   0.0005
        Signed Rank    S        39    Pr >= |S|   0.0005
```

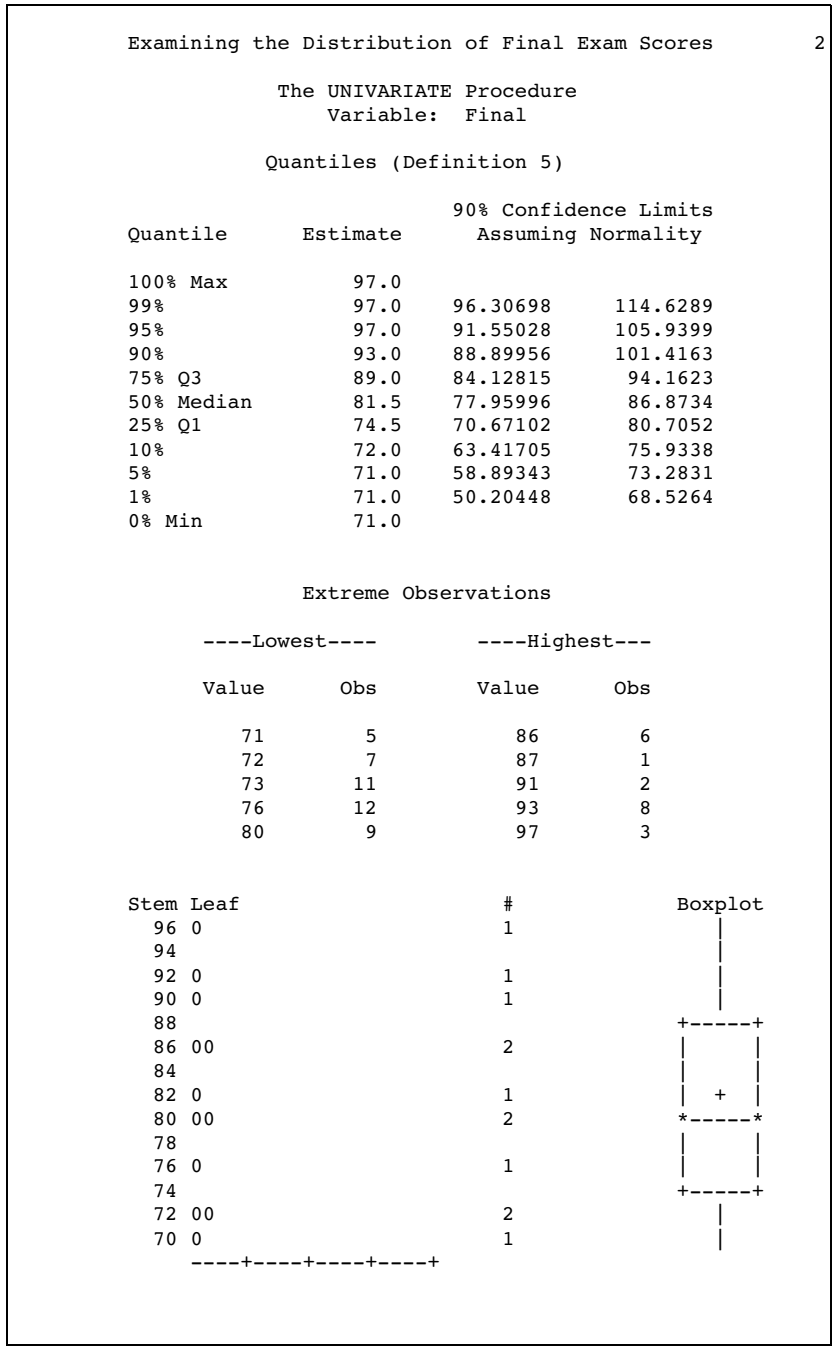


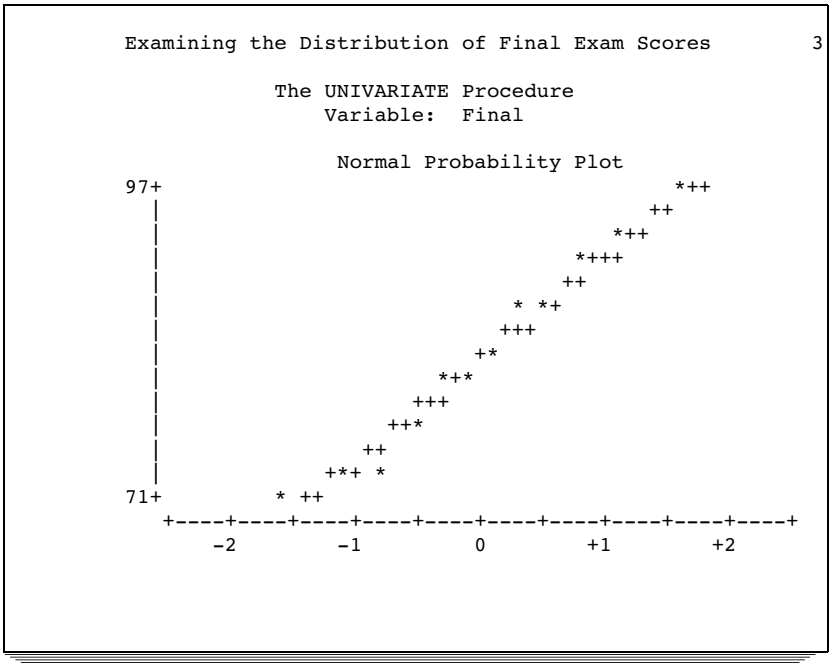
```
                    The SAS System                          2

                The UNIVARIATE Procedure
                    Variable:  Test1

                Quantiles (Definition 5)

                 Quantile      Estimate

                 100% Max          95.0
                 99%               95.0
                 95%               95.0
                 90%               94.0
                 75% Q3            90.5
                 50% Median        79.5
                 25% Q1            73.0
                 10%               63.0
                 5%                51.0
                 1%                51.0
                 0% Min            51.0


                  Extreme Observations

              ----Lowest----        ----Highest---

              Value      Obs        Value      Obs

                 51        2           87       12
                 63        4           89        8
                 71       10           92        6
                 75       11           94        1
                 75        7           95        3
```

Output 41.2 A Univariate Analysis with Tests for Normality and Plots of the Data Distribution

```
        Examining the Distribution of Final Exam Scores        1

                    The UNIVARIATE Procedure
                        Variable:  Final

                             Moments

N                          12    Sum Weights                12
Mean                82.4166667    Sum Observations          989
Std Deviation       8.59659905    Variance            73.9015152
Skewness            0.22597472    Kurtosis            -1.0846549
Uncorrected SS           82323    Corrected SS        812.916667
Coeff Variation    10.4306561    Std Error Mean      2.48162439


                    Basic Statistical Measures

         Location                      Variability

     Mean     82.41667     Std Deviation            8.59660
     Median   81.50000     Variance                73.90152
     Mode     80.00000     Range                   26.00000
                           Interquartile Range     14.50000


           Basic Confidence Limits Assuming Normality

           Parameter          Estimate    Lower 90% CL

           Mean               82.41667      79.03314
           Std Deviation       8.59660       6.85984
           Variance           73.90152      47.05738


                    Tests for Location: Mu0=80

        Test           -Statistic-    -----p Value------

        Student's t    t  0.973825    Pr > |t|    0.3511
        Sign           M         1    Pr >= |M|   0.7539
        Signed Rank    S         8    Pr >= |S|   0.4434


                      Tests for Normality

  Test                  --Statistic---    ------p Value------

  Shapiro-Wilk          W     0.952903    Pr < W       0.6797
  Kolmogorov-Smirnov    D     0.113328    Pr > D      >0.1500
  Cramer-von Mises      W-Sq  0.028104    Pr > W-Sq   >0.2500
  Anderson-Darling      A-Sq  0.212693    Pr > A-Sq   >0.2500
```

```
Examining the Distribution of Final Exam Scores        2

            The UNIVARIATE Procedure
                Variable:  Final

            Quantiles (Definition 5)

                                 90% Confidence Limits
Quantile        Estimate          Assuming Normality

100% Max           97.0
99%                97.0      96.30698      114.6289
95%                97.0      91.55028      105.9399
90%                93.0      88.89956      101.4163
75% Q3             89.0      84.12815       94.1623
50% Median         81.5      77.95996       86.8734
25% Q1             74.5      70.67102       80.7052
10%                72.0      63.41705       75.9338
5%                 71.0      58.89343       73.2831
1%                 71.0      50.20448       68.5264
0% Min             71.0


                Extreme Observations

      ----Lowest----          ----Highest---

      Value      Obs          Value      Obs

         71        5             86        6
         72        7             87        1
         73       11             91        2
         76       12             93        8
         80        9             97        3


Stem Leaf                      #             Boxplot
  96 0                         1                |
  94                                            |
  92 0                         1                |
  90 0                         1                |
  88                                          +-----+
  86 00                        2              |     |
  84                                          |     |
  82 0                         1              |  +  |
  80 00                        2              *-----*
  78                                          |     |
  76 0                         1              |     |
  74                                          +-----+
  72 00                        2                |
  70 0                         1                |
     ----+----+----+----+
```





Output 41.3 An Output Data Set That Contains Univariate Statistics

Quantile Statistics for Final Exam Scores 4
Output Data Set from PROC UNIVARIATE

Median	Pctl_Top	Pctl_Mid	Pctl_Low	Pctl_70
81.5	97	81.5	73	87

Procedure Syntax

Tip: Supports the Output Delivery System. See "Output Delivery System" on page 19

Reminder: You can use the ATTRIB, FORMAT, LABEL, and WHERE statements. See Chapter 3, "Statements with the Same Function in Multiple Procedures," for details. You can also use any global statements as well. See Chapter 2, "Fundamental Concepts for Using Base SAS Procedures," for a list.

PROC UNIVARIATE <*option(s)*>;

BY <DESCENDING> *variable-1* <…<DESCENDING> *variable-n*> <NOTSORTED>;

CLASS *variable-1*<(*variable-option(s)*)> <*variable-2*<(*variable-option(s)*)>> </ **KEYLEVEL=**'*value1*'|('*value1*' '*value2*')>;

FREQ *variable*;

HISTOGRAM <*variable(s)*> </ *option(s)*>;

ID *variable(s)*;

INSET <*keyword(s)* DATA=*SAS-data-set*> </ *option(s)*>;

OUTPUT <OUT=*SAS-data-set*> *statistic-keyword-1=name(s)* <… *statistic-keyword-n=name(s)*> <*percentiles-specification*>;

PROBPLOT <*variable(s)*> </ *option(s)*>;
QQPLOT <*variable(s)*> </ *option(s)*>;
VAR *variable(s)*;
WEIGHT *variable*;

To do this	Use this statement
Calculate separate statistics for each BY group	BY
Specify up to two class variables to categorize the analysis	CLASS
Specify a variable that contains the frequency of each observation	FREQ
Create a high-resolution graph of a histogram	HISTOGRAM
Specify one or more variables whose values identify the extreme observations	ID
Inset a table of summary statistics in a high-resolution graph	INSET
Create an output data set that contains specified statistics	OUTPUT
Create a high-resolution graph of a probability plot	PROBPLOT
Create a high-resolution graph of a quantile-quantile plot	QQPLOT
Select the analysis variables and determine their order in the report	VAR
Identify a variable whose values weight each observation in the statistical calculations	WEIGHT

PROC UNIVARIATE Statement

PROC UNIVARIATE <*option(s)*>;

To do this	Use this option
Specify the input data set	DATA=
Specify the input data set that contains annotate variables	ANNOTATE=
Specify the SAS catalog to save high-resolution graphics output	GOUT=
Control the statistical analysis	
Request all statistics and tables that the FREQ, MODES, NEXTRVAL=, PLOT, and CIBASIC options generate	ALL
Specify the confidence level for the confidence limits	ALPHA=
Request confidence limits for the mean, standard deviation, and variance based on normally distributed data	CIBASIC
Request confidence limits for quantiles using a distribution-free method	CIPCTLDF
Request confidence limits for quantiles based on normally distributed data	CIPCTLNORMAL

To do this	Use this option
Exclude observations with nonpositive weights from the analysis	EXCLNPWGT
Specify the value of the mean or location parameter	MU0=
Specify the number of extreme observations displayed	NEXTROBS=
Specify the number of extreme values displayed	NEXTRVAL=
Request tests for normality	NORMAL
Specify the mathematical definition used to compute quantiles	PCTLDEF=
Compute robust estimates of scale	ROBUSTSCALE
Specify the units to round the analysis variable prior to computing statistics	ROUND=
Compute trimmed means	TRIMMED=
Specify the variance divisor	VARDEF=
Compute Winsorized means	WINSORIZED=
Control the displayed output	
Request a frequency table	FREQ
Request a table that shows number of observations greater than, equal to, and less than MU0=	LOCCOUNT
Request a table of all possible modes	MODES
Suppress side-by-side plots	NOBYPLOT
Suppress tables of descriptive statistics	NOPRINT
Create low-resolution stem-and-leaf, box, and normal probability plots	PLOTS
Specify the approximate number of rows the plots use	PLOTSIZE=

Options

ALL
requests all statistics and tables that the FREQ, MODES, NEXTRVAL=5, PLOT, and CIBASIC options generate. If the analysis variables are not weighted, this option also requests the statistics and tables that the CIPCTLDF, CIPCTLNORMAL, LOCCOUNT, NORMAL, ROBUSTCALE, TRIMMED=.25, and WINSORIZED=.25 options generate. PROC UNIVARIATE also uses any values that you specify for ALPHA=, MU0=, NEXTRVAL=, CIBASIC, CIPCTLDF, CIPCTLNORMAL, TRIMMED=, or WINSORIZED= to produce the output.

ALPHA=*value*
specifies the default confidence level to compute confidence limits. The percentage for the confidence limits is (1–*value*) × 100. For example, ALPHA=.05 results in a 95 percent confidence limit.

Default: .05

Range: between 0 and 1

Main discussion: "Confidence Limits for Parameters of the Normal Distribution" on page 1393

Featured in: Example 4 on page 1428 and Example 5 on page 1432

ANNOTATE=*SAS-data-set*
specifies an input data set that contains annotate variables as described in *SAS/GRAPH Software: Reference*. You can use this data set to add features to your high-resolution graphics. PROC UNIVARIATE adds the features in this data set to every high-resolution graph that is produced in the PROC step.

Alias: ANNO=

Interaction: PROC UNIVARIATE does not use the ANNOTATE= data set unless you create a high-resolution graph with the HISTOGRAM, PROBPLOT, or QQPLOT statement.

Tip: Use the ANNOTATE= option in the HISTOGRAM, PROBPLOT, or QQPLOT statement if you want to add a feature to a specific graphics display.

CIBASIC<(<TYPE=*keyword*> <ALPHA=*value*>)>
requests confidence limits for the mean, standard deviation, and variance based on the assumption that the data are normally distributed. For large sample sizes, this assumption is not required for the mean because of the Central Limit Theorem.

TYPE=*keyword*
specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED.

Default: TWOSIDED

ALPHA=*value*
specifies the confidence level to compute the confidence limit. The percentage for the confidence limits is (1–*value*) × 100. For example, ALPHA=.05 results in a 95 percent confidence limit.

Default: The value of ALPHA= in the PROC statement

Range: between 0 and 1

Requirement: You must use the default value of VARDEF=, which is DF.

Main discussion: "Confidence Limits for Parameters of the Normal Distribution" on page 1393

Featured in: Example 4 on page 1428 and Example 5 on page 1432

CIPCTLDF<(<TYPE=*keyword*> <ALPHA=*value*>)>
requests confidence limits for quantiles by using a method that is distribution-free. In other words, no specific parametric distribution such as the normal is assumed for the data. PROC UNIVARIATE uses order statistics (ranks) to compute the confidence limits as described by Hahn and Meeker (1991).

TYPE=*keyword*
specifies the type of confidence limit, where *keyword* is LOWER, UPPER, SYMMETRIC, or ASYMMETRIC.

Default: SYMMETRIC

ALPHA=*value*
specifies the confidence level to compute the confidence limit. The percentage for the confidence limits is (1–*value*) × 100. For example, ALPHA=.05 results in a 95 percent confidence limit.

Default: The value of ALPHA= in the PROC statement

Range: between 0 and 1

Alias: CIQUANTDF

Restriction: This option is not available if you specify a WEIGHT statement.

Main discussion: "Confidence Limits for Quantiles" on page 1404

Featured in: Example 4 on page 1428

CIPCTLNORMAL <(<TYPE=*keyword*> <ALPHA=*value*>)>
requests confidence limits for quantiles based on the assumption that the data are normally distributed.

TYPE=*keyword*
specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED.

Default: TWOSIDED

ALPHA=*value*
specifies the confidence level to compute the confidence limit. The percentage for the confidence limits is (1–*value*) × 100. For example, ALPHA=.05 results in a 95 percent confidence limit.

Default: The value of ALPHA= in the PROC statement

Range: between 0 and 1

Alias: CIQUANTNORMAL

Requirement: You must use the default value of VARDEF=, which is DF.

Restriction: This option is not available if you specify a WEIGHT statement.

Main discussion: "Confidence Limits for Quantiles" on page 1404

Featured in: Example 5 on page 1432

DATA=*SAS-data-set*
specifies the input SAS data set.

Main discussion: "Input Data Sets" on page 18

EXCLNPWGT
excludes observations with nonpositive weight values (zero or negative) from the analysis. By default, PROC UNIVARIATE treats observations with negative weights like those with zero weights and counts them in the total number of observations.

Requirement: You must use a WEIGHT statement.

See also: "WEIGHT Statement" on page 1387

FREQ
requests a frequency table that consists of the variable values, frequencies, cell percentages, and cumulative percentages.

Interaction: If you specify the WEIGHT statement, PROC UNIVARIATE includes the weighted count in the table and uses this value to compute the percentages.

Featured in: Example 2 on page 1421

GOUT=*graphics-catalog*
specifies the SAS catalog that PROC UNIVARIATE uses to save the high-resolution graphics output.

Tip: If you omit the libref, PROC UNIVARIATE looks for the catalog in the temporary library called WORK and creates the catalog if it does not exist.

See also: For information on storing graphics output in SAS catalogs, see *SAS/GRAPH Software: Reference*

LOCCOUNT
requests a table that shows the number of observations greater than, equal to, and less than the value of MU0=. PROC UNIVARIATE uses these values to construct the sign test and the signed rank test.

Restriction: This option is not available if you specify a WEIGHT statement.

See also: MU0= on page 1329

Featured in: Example 4 on page 1428

MODES
requests a table of all possible modes. By default, when the data contain multiple modes, PROC UNIVARIATE displays the lowest mode in the table of basic statistical measures. When all the values are unique, PROC UNIVARIATE does not produce a table of modes.

Alias: MODE

Main discussion: "Calculating the Mode" on page 1406

Featured in: Example 4 on page 1428

MU0=*value(s)*
specifies the value of the mean or location parameter (μ_0) in the null hypothesis for tests of location. If you specify one value, PROC UNIVARIATE tests the same null hypothesis for all analysis variables. If you specify multiple values, a VAR statement is required, and PROC UNIVARIATE tests a different null hypothesis for each analysis variable in the corresponding order.

Alias: LOCATION=

Default: 0

Main discussion: "Tests for Location" on page 1394

Example: The following statement tests if the mean of the first variable equals 0 and the mean of the second variable equals 0.5.

```
proc univariate mu0=0 0.5;
```

Featured in: Example 5 on page 1432

NEXTROBS=*n*
specifies the number of extreme observations that PROC UNIVARIATE lists in the table of extreme observations. The table lists the *n* lowest observations and the *n* highest observations.

Default: 5

Range: an integer between 0 and the half the maximum number of observations

Tip: Use NEXTROBS=0 to suppress the table of extreme observations.

Featured in: Example 2 on page 1421 and Example 7 on page 1439

NEXTRVAL=*n*
specifies the number of extreme values that PROC UNIVARIATE lists in the table of extreme values. The table lists the *n* lowest unique values and the *n* highest unique values.

Default: 0

Range: an integer between 0 and half the maximum number of observations

Featured in: Example 2 on page 1421

NOBYPLOT
suppresses side-by-side box plots when you use the BY statement and the ALL option or the PLOT option in the PROC statement.

NOPRINT
suppresses all the tables of descriptive statistics that the PROC UNIVARIATE statement creates. NOPRINT does not suppress the tables that the HISTOGRAM statement creates.

Tip: Use NOPRINT when you want to create an OUT= output data set only.

Featured in: Example 6 on page 1437 and Example 8 on page 1444

NORMAL
requests tests for normality that include the Shapiro-Wilk test and a series of goodness-of-fit tests based on the empirical distribution function.

Alias: NORMALTEST

Restriction: This option is not available if you specify a WEIGHT statement.

Main discussion: "Goodness-of-Fit Tests" on page 1396

Featured in: Example 5 on page 1432

PCTLDEF=*value*

specifies the definition that PROC UNIVARIATE uses to calculate quantiles.

Alias: DEF=

Default: 5

Range: 1, 2, 3, 4, 5

Restriction: You cannot use PCTLDEF= when you compute weighted quantiles.

Main discussion:

PLOTS

produces a stem-and-leaf plot (or a horizontal bar chart), a box plot, and a normal probability plot. If you use a BY statement, side-by-side box plots that are labeled **`Schematic Plots`** appear after the univariate analysis for the last BY group.

Alias: PLOT

Main discussion: "Generating Line Printer Plots" on page 1389

Featured in: Example 5 on page 1432 and Example 7 on page 1439

PLOTSIZE=*n*

specifies the approximate number of rows that the plots use. If *n* is larger than the value of the SAS system option PAGESIZE=, PROC UNIVARIATE uses the value of PAGESIZE=. If *n* is less than eight, PROC UNIVARIATE uses eight rows to draw the plots.

Default: the value of PAGESIZE=

Range: 8 to the value of PAGESIZE=

Featured in: Example 5 on page 1432 and Example 7 on page 1439

ROBUSTSCALE

produces a table with robust estimates of scale. The statistics include the interquartile range, Gini's mean difference, the median absolute deviation about the median (*MAD*), and two statistics proposed by Rousseeuw and Croux (1993), Q_n, and S_n.

Restriction: This option is not available if you specify a WEIGHT statement.

Main discussion: "Robust Measures of Scale" on page 1403

Featured in: Example 3 on page 1424

ROUND=*unit(s)*

specifies the units to use to round the analysis variables prior to computing statistics. If you specify one unit, PROC UNIVARIATE uses this unit to round all analysis variables. If you specify multiple units, a VAR statement is required, and each unit rounds the values of the corresponding analysis variable. If ROUND=0, no rounding occurs.

Default: 0

Tip: ROUND= reduces the number of unique variable values, thereby reducing the memory requirements.

Range: ≥ 0

Main discussion: "Rounding" on page 1388

Example: To make 1 the rounding unit for the first analysis variable and 0.5 the rounding unit for second analysis variable, submit the statement

```
proc univariate round=1 0.5;
```

Featured in: Example 2 on page 1421

TRIMMED=*value(s)* <(<TYPE=*keyword*> <ALPHA=*value*>)>
requests a table of trimmed means, where *value* specifies the number or the proportion of observations that PROC UNIVARIATE trims. If *value* is a proportion p between 0 and .5, the number of observations that PROC UNIVARIATE trims is the smallest integer that is greater than or equal to np, where n is the number of observations.

TYPE=*keyword*
specifies the type of confidence limit for the mean, where *keyword* is LOWER, UPPER, or TWOSIDED.

Default: TWOSIDED

ALPHA=*value*
specifies the confidence level to compute the confidence limit. The percentage for the confidence limits is (1–*value*) × 100. For example, ALPHA=.05 results in a 95 percent confidence limit.

Default: The value of ALPHA= in the PROC statement

Range: between 0 and 1

Alias: TRIM=

Range: between 0 and half the number of nonmissing observations. When a proportion is specified, *value* must be less than .5.

Requirement: To compute confidence limits for the mean and the Student's t test, you must use the default value of VARDEF=, which is DF.

Restriction: This option is not available if you specify a WEIGHT statement.

Main discussion "Trimmed Means" on page 1402

Featured in: Example 3 on page 1424

VARDEF=*divisor*
specifies the divisor to use in the calculation of variances and standard deviation. Table 41.1 on page 1331 shows the possible values for *divisor* and associated divisors.

Table 41.1 Possible Values for VARDEF=

Value	Divisor	Formula for Divisor
DF	degrees of freedom	$n - 1$
N	number of observations	n
WDF	sum of weights minus one	$(\Sigma_i\, w_i) - 1$
WEIGHT | WGT	sum of weights	$\Sigma_i\, w_i$

The procedure computes the variance as $CSS/divisor$, where CSS is the corrected sums of squares and equals $\sum (x_i - \overline{x})^2$. When you weight the analysis variables, CSS equals $\sum w_i (x_i - \overline{x}_w)^2$, where $\overline{x}_w$ is the weighted mean.

Default: DF

Requirement: To compute the standard error of the mean, confidence limits, and Student's t test, use the default value of VARDEF=.

Tip: When you use the WEIGHT statement and VARDEF=DF, the variance is an estimate of σ^2, where the variance of the *i*th observation is $var\left(x_i\right) = \sigma^2/w_i$ and w_i is the weight for the *i*th observation. This yields an estimate of the variance of an observation with unit weight.

Tip: When you use the WEIGHT statement and VARDEF=WGT, the computed variance is asymptotically (for large n) an estimate of $\sigma^2/\overline{w}$, where $\overline{w}$ is the average weight. This yields an asymptotic estimate of the variance of an observation with average weight.

See also: "Keywords and Formulas" on page 1458 and "WEIGHT Statement" on page 1387

WINSORIZED=*value(s)* <(<TYPE=*keyword*> <ALPHA=*value*>)>
requests of a table of Winsorized means, where *value* is the number or the proportion of observations that PROC UNIVARIATE uses to compute the Winsorized mean. If *value* is a proportion *p* between 0 and .5, the number of observations that PROC UNIVARIATE uses is equal to the smallest integer that is greater than or equal to *np*, where *n* is the number of observations.

TYPE=*keyword*
specifies the type of confidence limit for the mean, where *keyword* is LOWER, UPPER, or TWOSIDED.

Default: TWOSIDED

ALPHA=*value*
specifies the confidence level to compute the confidence limit. The percentage for the confidence limits is (1–*value*) × 100. For example, ALPHA=.05 results in a 95 percent confidence limit.

Default: The value of ALPHA= in the PROC statement

Range: between 0 and 1

Alias: WINSOR=

Range: between 0 and half the number of nonmissing observations. When a proportion is specified, *value* must be less than .5.

Requirement: To compute confidence limits and the Student's *t* test, you must use the default value of VARDEF=, which is DF.

Restriction: This option is not available if you specify a WEIGHT statement.

Main discussion "Winsorized Means" on page 1401

Featured in: Example 3 on page 1424

BY Statement

Calculates univariate statistics separately for each BY group.

Main discussion: "BY" on page 68

Featured in: Example 7 on page 1439

BY <DESCENDING> *variable-1* <…<DESCENDING> *variable-n*><NOTSORTED>;

Required Arguments

variable
specifies the variable that the procedure uses to form BY groups. You can specify more than one variable. If you do not use the NOTSORTED option in the BY statement, the observations in the data set must either be sorted by all the variables that you specify, or they must be indexed appropriately. These variables are called *BY variables*.

Options

DESCENDING
specifies that the data set is sorted in descending order by the variable that immediately follows the word DESCENDING in the BY statement.

NOTSORTED
specifies that observations are not necessarily sorted in alphabetic or numeric order. The data are grouped in another way, for example, chronological order.

The requirement for ordering or indexing observations according to the values of BY variables is suspended for BY-group processing when you use the NOTSORTED option. In fact, the procedure does not use an index if you specify NOTSORTED. The procedure defines a BY group as a set of contiguous observations that have the same values for all BY variables. If observations with the same values for the BY variables are not contiguous, the procedure treats each contiguous set as a separate BY group.

CLASS Statement

Specifies up to two variables whose values define the classification levels for the analysis.

Interaction: When you use the HISTOGRAM, PROBPLOT, or QQPLOT statement, PROC UNIVARIATE creates comparative histograms, comparative probability plots, or comparative quantile-quantile plots.

Featured in: Example 10 on page 1450

CLASS *variable-1*<(*variable-option(s)*)> <*variable-2*<(*variable-option(s)*)>>
</ **KEYLEVEL=**'*value1*'|('*value1*' '*value2*')>;

Required Arguments

variable-n
specifies one or two variables that the procedure uses to group the data into classification levels. Variables in a CLASS statement are referred to as *class variables*.

Class variables can be numeric or character. Class variables can have continuous values, but they typically have a few discrete values that define levels of the variable.

You do not have to sort the data by class variables. PROC UNIVARIATE uses the formatted values of the class variables to determine the classification levels.

You can use the HISTOGRAM, PROBPLOT, or QQPLOT statement with the CLASS statement to create one-way and two-way comparative plots. When you use one class variable, PROC UNIVARIATE displays an array of component plots (stacked or side-by-side), one for each level of the classification variable. When you use two class variables, PROC UNIVARIATE displays a matrix of component plots, one for each combination of levels of the classification variables. The observations in a given level are referred to collectively as a *cell*.

Restriction: The length of a character class variable cannot exceed 16.

Interaction: When you create a one-way comparative plot, the observations in the input data set are sorted by the formatted values (levels) of the variable. PROC UNIVARIATE creates a separate plot for the analysis variable values in each level, and arranges these component plots in an array to form the comparative plot with uniform horizontal and vertical axes.

When you create a two-way comparative plot, the observations in the input data set are cross-classified according to the values (levels) of these variables. PROC UNIVARIATE creates a separate plot for the analysis variable values in each cell of the cross-classification and arranges these component plots in a matrix to form the comparative plot with uniform horizontal and vertical axes. The levels of *variable–1* are the labels for the rows of the matrix, and the levels of *variable–2* are the labels for the columns of the matrix.

Interaction: If you associate a label with a variable, PROC UNIVARIATE displays the variable label in the comparative plot and this label is parallel to the column (or row) labels.

Tip: Use the MISSING option to treat missing values as valid levels.

Tip: To reduce the number of classification levels, use a FORMAT statement to combine variable values.

Options

KEYLEVEL='*value1*'|('*value1*' '*value2*')

specifies the *key cell* in a comparative plot. PROC UNIVARIATE first determines the bin size and midpoints for the key cell, and then extends the midpoint list to accommodate the data ranges for the remaining cells. Thus, the choice of the key cell determines the uniform horizontal axis that PROC UNIVARIATE uses for all cells.

If you specify only one class variable and use a HISTOGRAM statement, KEYLEVEL='*value*' identifies the key cell as the level for which *variable* is equal to *value*. By default, PROC UNIVARIATE sorts the levels in the order that is determined by the ORDER= option. Then, the key cell is the first occurrence of a level in this order. The cells display in order from top to bottom or left to right. Consequently, the key cell appears at the top (or left). When you specify a different key cell with the KEYLEVEL= option, this cell appears at the top (or left).

Likewise, with the PROBPLOT statement and the QQPLOT statement the key cell determines uniform axis scaling.

If you specify two class variables, use KEYLEVEL=('*value1*' '*value2*') to identify the key cell as the level for which *variable-n* is equal to *value-n*. By default, PROC UNIVARIATE sorts the levels of the first variable in the order that is determined by its ORDER= option and, within each of these levels, it sorts the levels of the second variable in the order that is determined by its ORDER= option. Then, the default key cell is the first occurrence of a combination of levels for the two variables in this order. The cells display in the order of *variable–1* from top to bottom and in the order

of *variable–2* from left to right. Consequently, the default key cell appears at the upper left corner. When you specify a different key cell with the KEYLEVEL= option, this cell appears at the upper left corner.

Restriction: The length of the KEYLEVEL= value cannot exceed 16 characters and you must specify a formatted value.

Requirement: This option is ignored unless you specify a HISTOGRAM, PROBPLOT, or QQPLOT statement.

See also: the ORDER= option on page 1335

MISSING
specifies to treat the missing values for the class variable as valid classification levels. Special missing values that represent numeric values (the letters A through Z and the underscore (_) character) are each considered as a separate value.

Default: If you omit MISSING, PROC UNIVARIATE excludes the observations with a missing class variable value from the analysis.

Requirement: Enclose this option in parentheses after the class variable.

See also: *SAS Language Reference: Concepts* for a discussion of missing values that have special meaning.

ORDER=DATA | FORMATTED | FREQ | INTERNAL
specifies the display order for the class variable values, where

DATA
orders values according to their order in the input data set.

Interaction: When you use a HISTOGRAM, PROBPLOT, or QQPLOT statement, PROC UNIVARIATE displays the rows (columns) of the comparative plot from top to bottom (left to right) in the order that the class variable values first appear in the input data set.

FORMATTED
orders values by their ascending formatted values. This order depends on your operating environment.

Interaction: When you use a HISTOGRAM, PROBPLOT, or QQPLOT statement, PROC UNIVARIATE displays the rows (columns) of the comparative plot from top to bottom (left to right) in increasing order of the formatted class variable values. For example, a numeric class variable DAY (with values 1, 2, and 3) has a user-defined format that assigns **`Wednesday`** to the value 1, **`Thursday`** to the value 2, and **`Friday`** to the value 3. The rows of the comparative plot will appear in alphabetical order (Friday, Thursday, Wednesday) from top to bottom.

FREQ
orders values by descending frequency count so that levels with the most observations are listed first. If two or more values have the same frequency count, PROC UNIVARIATE uses the formatted values to determine the order.

Interaction: When you use a HISTOGRAM, PROBPLOT, or QQPLOT statement, PROC UNIVARIATE displays the rows (columns) of the comparative plot from top to bottom (left to right) in order of decreasing frequency count for the class variable values.

INTERNAL
orders values by their unformatted values, which yields the same order as PROC SORT. This order depends on your operating environment.

If there are two or more distinct internal values with the same formatted value then PROC UNIVARIATE determines the order by the internal value that occurs first in the input data set.

Interaction: When you use a HISTOGRAM, PROBPLOT, or QQPLOT statement, PROC UNIVARIATE displays the rows (columns) of the comparative plot from top to bottom (left to right) in increasing order of the internal (unformatted) values of the class variable. The first class variable is used to label the rows of the comparative plots (top to bottom). The second class variable are used to label the columns of the comparative plots (left to right). For example, a numeric class variable DAY (with values 1, 2, and 3) has a user-defined format that assigns **`Wednesday`** to the value 1, **`Thursday`** to the value 2, and **`Friday`** to the value 3. The rows of the comparative plot will appear in day-of-the-week order (Wednesday, Thursday, Friday) from top to bottom.

Default: INTERNAL

Requirement: Enclose this option in parentheses after the class variable.

Interaction: When you use a HISTOGRAM, PROBPLOT, or QQPLOT statement and ORDER=INTERNAL, PROC UNIVARIATE constructs the levels of the class variables by using the formatted values of the variables. The formatted values of the first class variable are used to label the rows of the comparative plots (top to bottom). The formatted values of a second class variable are used to label the columns of the comparative plots (left to right).

PROC UNIVARIATE determines the layout of a two-way comparative plot by using the order for the first class variable to obtain the order of the rows from top to bottom. Then it applies the order for the second class variable to the observations that correspond to the first row to obtain the order of the columns from left to right. If any columns remain unordered (that is, the categories are unbalanced), PROC UNIVARIATE applies the order for the second class variable to the observations in the second row, and so on, until all the columns have been ordered.

Featured in: Example 10 on page 1450

FREQ Statement

Specifies a numeric variable whose values represent the frequency of the observation.

Tip: The FREQ statement affects the degrees of freedom, but the WEIGHT statement does not.

See also: For an example that uses the FREQ statement, see “FREQ” on page 70

FREQ *variable*;

Required Arguments

variable
specifies a numeric variable whose value represents the frequency of the observation. If you use the FREQ statement, the procedure assumes that each observation represents *n* observations, where *n* is the value of *variable*. If *variable* is not an integer, the SAS System truncates it. If *variable* is less than 1 or is missing, the procedure excludes that observation from the analysis.

HISTOGRAM Statement

Creates histograms using high-resolution graphics and optionally superimposes parametric and nonparametric density curve estimates.

Alias: HIST

Tip: You can use multiple HISTOGRAM statements.

Featured in: Example 8 on page 1444 and Example 10 on page 1450

HISTOGRAM <*variable(s)*> </ *option(s)*>;

To do this	Use this option
Create output data set with information on histogram intervals	OUTHISTOGRAM=
Request estimated density curve	
Fit beta density with threshold parameter θ, scale parameter σ, and shape parameters α and β	BETA(*beta-suboptions*)
Fit exponential density with threshold parameter θ and scale parameter σ	EXPONENTIAL(*exponential-suboptions*)
Fit gamma density with threshold parameter θ, scale parameter σ, and shape parameter α	GAMMA(*gamma-suboptions*)
Fit nonparametric kernel density estimates	KERNEL(*kernel-suboptions*)
Fit lognormal density with threshold parameter θ, scale parameter ζ, and shape parameter σ	LOGNORMAL(*lognormal-suboptions*)
Fit normal density with mean μ and standard deviation σ	NORMAL(*normal-suboptions*)
Fit Weibull density with threshold parameter θ, scale parameter σ, and shape parameter c	WEIBULL(*Weibull-suboptions*)
Parametric density curve suboptions	
Specify shape parameter α for fitted beta or gamma curve	ALPHA=
Specify second shape parameter β for beta fitted curve	BETA=
Specify shape parameter c for fitted Weibull curve	C=
Specify the mean μ for fitted normal curve	MU=
Specify scale parameter σ for the fitted beta curve, exponential curve, gamma curve and Weibull curve; standard deviation σ for fitted normal curve; or the scale parameter σ for the fitted lognormal curve	SIGMA=
Specify threshold parameter θ for fitted beta curve, exponential curve, gamma curve, lognormal curve, and Weibull curve	THETA=

To do this	Use this option
Specify scale parameter ζ for fitted lognormal curve	ZETA=
Nonparametric density curve suboptions	
Specify standardized bandwidth parameter c for fitted kernel density estimates	C=
Specify type of kernel density curve	K=
Control appearance of fitted density curves	
Specify color of fitted curve	COLOR=
Fill area under fitted curve	FILL
Specify line type of fitted curve	L=
Display table of histogram interval midpoints	MIDPERCENTS
Suppress the table summarizing the fitted curve	NOPRINT
List percentages for calculated and estimated quantiles	PERCENTS=
Specify width of fitted density curve	W=
Control general histogram layout	
Specify width for the bars	BARWIDTH=
Force creation of a histogram	FORCEHIST
Create a grid	GRID
Specify offset for horizontal axis	HOFFSET=
Specify reference lines perpendicular to the horizontal axis	HREF=
Specify labels for HREF= lines	HREFLABELS=
Specify vertical position of labels for HREF= lines	HREFLABPOS=
Specify a line style for grid lines	LGRID=
List percentages for histogram intervals	MIDPOINTS=
Suppress histogram bars	NOBARS
Suppress frame around plotting area	NOFRAME
Suppress label for horizontal axis	NOHLABEL
Suppress plot	NOPLOT
Suppress label for vertical axis	NOVLABEL
Suppress tick marks and tick mark labels for vertical axis	NOVTICK
Include right endpoint in interval	RTINCLUDE
Turn and vertically string out characters in labels for vertical axis	TURNVLABELS
Specify tick mark values for vertical axis	VAXIS=
Specify label for vertical axis	VAXISLABEL=
Specify length of offset at upper end of vertical axis	VOFFSET=

To do this	Use this option
Specify reference lines perpendicular to the vertical axis	VREF=
Specify labels for VREF= lines	VREFLABELS=
Specify horizontal position of labels for VREF= lines	VREFLABPOS=
Specify scale for vertical axis	VSCALE=
Specify line thickness for axes and frame	WAXIS=
Specify line thickness for grid	WGRID=
Enhance the graph	
Specify annotate data set	ANNOTATE=
Specify color for axis	CAXIS=
Specify color of outlines of histogram bars	CBARLINE=
Specify color for filling under curve	CFILL=
Specify color for frame	CFRAME=
Specify color for grid lines	CGRID=
Specify color for HREF= lines	CHREF=
Specify color for text	CTEXT=
Specify color for VREF= lines	CVREF=
Specify description for plot in graphics catalog	DESCRIPTION=
Specify software font for text	FONT=
Specify height of text used outside framed areas	HEIGHT=
Specify number of horizontal minor tick marks	HMINOR=
Specify software font for text inside framed areas	INFONT=
Specify height of text inside framed areas	INHEIGHT=
Specify line style for HREF= lines	LHREF=
Specify line style for VREF= lines	LVREF=
Specify name for plot in graphics catalog	NAME=
Specify pattern for filling under curve	PFILL=
Specify number of vertical minor tick marks	VMINOR=
Specify line thickness for bar outlines	WBARLINE=
Enhance comparative histograms	
Apply annotation requested in ANNOTATE= data set to key cell only	ANNOKEY
Specify color for filling frame for row labels	CFRAMESIDE=
Specify color for filling frame for column labels	CFRAMETOP=
Specify color for proportion of frequency bar	CPROP=
Specify distance between tiles	INTERTILE=
Specify maximum number of bins to display	MAXNBIN=

To do this	Use this option
Limit the number of bins that display to within a specified number of standard deviations above and below mean of data in key cell	MAXSIGMAS=
Specify number of columns in comparative histogram	NCOLS=
Specify number of rows in comparative histogram	NROWS=

Arguments

variable(s)
identifies one or more analysis variables that the procedure uses to create histograms.

Default: If you omit *variable(s)* in the HISTOGRAM statement, then the procedure creates a histogram for each variable that you list in the VAR statement, or for each numeric variable in the DATA= data set if you omit a VAR statement.

Requirement: If you specify a VAR statement, use a subset of the *variable(s)* that you list in the VAR statement. Otherwise, *variable(s)* are any numeric variables in the DATA= data set.

Options

ALPHA=*value*
specifies the shape parameter α for fitted density curves when you request the BETA and GAMMA options.

Alias: A= if you use it as a *beta-suboption*. SHAPE= if you use it as a *gamma-suboption*

Default: a maximum likelihood estimate

Requirement: Enclose this suboption in parentheses after the BETA option or GAMMA option.

ANNOKEY
specifies to apply the annotation requested with the ANNOTATE= option to the *key cell* only. By default, PROC UNIVARIATE applies annotation to all of the cells.

Requirement: This option is ignored unless you specify the CLASS statement.

Tip: Use the KEYLEVEL= option in the CLASS statement to specify the key cell.

See also: the KEYLEVEL= option on page 1334

ANNOTATE=*SAS-data-set*
specifies an input data set that contains annotate variables as described in *SAS/GRAPH Software: Reference*.

Alias: ANNO=

Tip: You can also specify an ANNOTATE= data set in the PROC UNIVARIATE statement to enhance all the graphic displays that the procedure creates.

See also: ANNOTATE= on page 1327 in the PROC UNIVARIATE statement

BARWIDTH=*value*
specifies the width of the histogram bars in screen percent units.

BETA<(*beta-suboptions*)>
displays a fitted beta density curve on the histogram.

Restriction: The BETA option can occur only once in a HISTOGRAM statement.

Interaction: The beta distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. Use the THETA= and SIGMA= suboptions to specify these parameters. The default values for THETA= and SIGMA= are 0 and 1, respectively. You can specify THETA=EST and SIGMA=EST to request maximum likelihood estimates for θ and σ.

Note: Three- and four-parameter maximum likelihood estimation may not always converge. △

Interaction: The beta distribution has two shape parameters, α and β. If these parameters are known, you can specify their values with the ALPHA= and BETA= options. By default, PROC UNIVARIATE computes maximum likelihood estimates for α and β.

Main Discussion: See "Beta Distribution" on page 1406

See also: the ALPHA= suboption on page 1340, BETA= suboption on page 1341, SIGMA= suboption on page 1350, and THETA= suboption on page 1350

BETA=*value*

specifies the second shape parameter β for the fitted beta density curves when you request the BETA option.

Alias: B=

Default: a maximum likelihood estimate

Requirement: Enclose this suboption in parentheses after the BETA option.

C=*value*

specifies the shape parameter c for the fitted Weibull density curve when you request the WEIBULL option.

Default: a maximum likelihood estimate

Requirement: Enclose this suboption in parentheses after the WEIBULL option.

C=*value(s)*|MISE

specifies the standardized bandwidth parameter c for kernel density estimates when you request the KERNEL option.

Default: the bandwidth that minimizes the approximate MISE.

Restriction: You can specify up to five values to request multiple estimates.

Requirement: Enclose this suboption in parentheses after the KERNEL option.

Interaction: You can also use the C= suboption with the K= suboption, which specifies the kernel function, to compute multiple estimates. If you specify more kernel functions than bandwidths, PROC UNIVARIATE repeats the last bandwidth in the list for the remaining estimates. Likewise, if you specify more bandwidths than kernel functions, then PROC UNIVARIATE repeats the last kernel function for the remaining estimates. For example, the following statements compute three density estimates:

```
proc univariate;
   var length;
   histogram length / kernel(c=1 2 3 k=normal quadratic);
   run;
```

The first uses a normal kernel and a bandwidth of 1, the second uses a quadratic kernel and a bandwidth of 2, and the third uses a quadratic kernel and a bandwidth of 3.

Tip: To estimate a bandwidth that minimizes the approximate mean integrated square error (MISE) use the C=MISE suboption. For example, the following statements compute three density estimates:

```
proc univariate;
   var length;
   histogram length / kernel(c=0.5 1.0 mise);
  run;
```

The first two estimates have standardized bandwidths of 0.5 and 1.0, respectively, and the third has a bandwidth that minimizes the approximate MISE.

CAXIS=*color*
specifies the color for the axes and tick marks.

Alias: CAXES= and CA=

Default: the first color in the device color list

CBARLINE=*color*
specifies the color for the outline of the histogram bars.

Default: the first color in the device color list

Featured in: Example 8 on page 1444

CFILL=*color*
specifies the color to fill the bars of the histogram (or the area under a fitted density curve if you also specify the FILL option).

See also: FILL option on page 1343 and PFILL=option on page 1350

Featured in: Example 8 on page 1444 and Example 10 on page 1450

CFRAME=*color*
specifies the color for the area that is enclosed by the axes and frame.

Alias: CRF=

Default: The area is not filled.

CFRAMESIDE=*color*
specifies the color to fill the frame area for the row labels that display along the left side of the comparative histogram. This color also fills the frame area for the label of the corresponding class variable (if you associate a label with the variable).

Default: These areas are not filled.

Requirement: This option is ignored unless you specify the CLASS statement.

CFRAMETOP=*color*
specifies the color to fill the frame area for the column labels that display across the top of the comparative histogram. This color also fills the frame area for the label of the corresponding class variable (if you associate a label with the variable).

Default: These areas are not filled.

Requirement: This option is ignored unless you specify the CLASS statement.

CGRID=*color*
specifies the color for grid lines when a grid displays on the histogram.

Default: the first color in the device color list

Interaction: This option automatically invokes the GRID= option.

CHREF=*color*
specifies the color for horizontal axis reference lines when you specify the HREF= option.

Default: the first color in the device color list

COLOR=*color*
specifies the color of the density curve.

Requirement: You must enclose this suboption in parentheses after the density curve option or the KERNEL option.

Interaction: You can specify as a KERNEL suboption a list of up to five colors in parentheses for multiple kernel density estimates. If there are more estimates than colors, the remaining estimates use the last color that you specify.

CPROP=*color* | EMPTY
specifies the color for a horizontal bar whose length (relative to the width of the tile) indicates the proportion of the total frequency that is represented by the corresponding cell in a comparative histogram.

Default: bars do not display

Requirement: This option is ignored unless you specify the CLASS statement.

Tip: Use the keyword EMPTY to display empty bars.

CTEXT=*color*
specifies the color for tick mark values and axis labels.

Alias: CT=

Default: The color that you specify for the CTEXT= option in the GOPTIONS statement. If you omit the GOPTIONS statement, the default is the first color in the device color list.

CVREF=*color*
specifies the color for the reference lines that you request with the VREF= option.

Alias: CV=

Default: the first color in the device color list

DESCRIPTION='*string*'
specifies a description, up to 40 characters long, that appears in the PROC GREPLAY master menu.

Alias: DES=

Default: the variable name

EXPONENTIAL<(*exponential-suboptions*)>
displays a fitted exponential density curve on the histogram.

Alias EXP

Restriction: The EXPONENTIAL option can occur only once in a HISTOGRAM statement.

Interaction: The parameter θ must be less than or equal to the minimum data value. Use the THETA= suboption to specify θ. The default value for θ is zero. Specify THETA=EST to request the maximum likelihood estimate for θ.

Interaction: Use the SIGMA= suboption to specify σ. By default, PROC UNIVARIATE computes a maximum likelihood estimate for σ. For example, the following statements fit an exponential curve with $\theta = 10$ and with a maximum likelihood estimate for σ:

```
proc univariate;
   var length;
   histogram / exponential(theta=10 l=2 color=red);
run;
```

Main discussion: See "Exponential Distribution" on page 1407

See also: the SIGMA= suboption on page 1350 and THETA= suboption on page 1350

Featured in: Example 8 on page 1444

FILL
fills areas under the fitted density curve or the kernel density estimate with colors and patterns.

Restriction: The FILL suboption can occur with only one fitted curve.

Requirement: Enclose the FILL suboption in parentheses after a density curve option or the KERNEL option.

Interaction: The CFILL= and PFILL= options specify the color and pattern for the area under the curve.

See also: For a list of available colors and patterns, see *SAS/GRAPH Software: Reference*

Featured in: Example 8 on page 1444

FONT=*font*
specifies a software font for the axis labels.

Default: hardware characters

Interaction: The FONT= *font* takes precedence over the FTEXT= *font* that you specify in the GOPTIONS statement.

FORCEHIST
forces PROC UNIVARIATE to create a histogram when there is only one unique observation. By default, if the standard deviation of the data is zero then PROC UNIVARIATE does not create a histogram.

GAMMA<(*gamma-suboptions*)>
displays a fitted gamma density curve on the histogram.

Restriction: The GAMMA option can occur only once in a HISTOGRAM statement.

Interaction: The parameter θ must be less than the minimum data value. Use the THETA= suboption to specify θ. The default value for θ is zero. Specify THETA=EST to request the maximum likelihood estimate for θ.

Interaction: Use the ALPHA= and the SIGMA= suboptions to specify the shape parameter α and the scale parameter σ. By default, PROC UNIVARIATE computes maximum likelihood estimates for α and σ. For example, the following statements fit a gamma curve with $\theta = 4$ and with a maximum likelihood estimate for α and σ:

```
proc univariate;
   var length;
   histogram length/ gamma(theta=4);
run;
```

PROC UNIVARIATE calculates the maximum likelihood estimate of α iteratively using the Newton-Raphson approximation.

Main discussion: See "Gamma Distribution" on page 1407

See also: the SIGMA= suboption on page 1350, ALPHA= suboption on page 1340, and the THETA= suboption on page 1350

GRID
specifies to display a grid on the histogram. Grid lines are horizontal lines that are positioned at major tick marks on the vertical axis.

See also: the CGRID= option on page 1342

HEIGHT=*value*
specifies the height in percentage screen units of text for axis labels, tick mark labels, and legends. This option takes precedence over the HTEXT= option in the GOPTIONS statement.

HMINOR=*n*
specifies the number of minor tick marks between each major tick mark on the horizontal axis. PROC UNIVARIATE does not label minor tick marks.

Alias: HM=

Default: 0

HOFFSET=*value*
specifies the offset in percentage screen units at both ends of the horizontal axis.

Tip: Use HOFFSET=0 to eliminate the default offset.

HREF=*value(s)*
draws reference lines that are perpendicular to the horizontal axis at the values that you specify.

See also: CHREF= option on page 1342 and LHREF= option on page 1346.

HREFLABELS='*label1*' ... '*labeln*'
specifies labels for the reference lines that you request with the HREF= option.

Alias: HREFLABEL= and HREFLAB=

Restriction: The number of labels must equal the number of reference lines. Labels can have up to 16 characters.

HREFLABPOS=*n*
specifies the vertical position of HREFLABELS= labels, where *n* is

1	positions the labels along the top of the histogram
2	staggers the labels from top to bottom
3	positions the labels along the bottom.

Default: 1

INFONT=*font*
specifies a software font to use for text inside the framed areas of the histogram. The INFONT= option takes precedence over the FTEXT= option in the GOPTIONS statement.

See also: For a list of fonts, see *SAS/GRAPH Software: Reference*.

INHEIGHT=*value*
specifies the height, in percentage screen units of text, to use inside the framed areas of the histogram.

Default: The height that you specify with the HEIGHT= option. If you do not specify the HEIGHT= option, the default height is the height that you specify with the HTEXT= option in the GOPTIONS statement.

INTERTILE=*value*
specifies the distance in horizontal percentage screen units between the framed areas, which are called *tiles*.

Default: .75 in percentage screen units.

Requirement: This option is ignored unless you specify the CLASS statement.

Featured in: Example 10 on page 1450

K=NORMAL | QUADRATIC | TRIANGULAR
specifies the kernel function (normal, quadratic, or triangular) that PROC UNIVARIATE uses to compute a kernel density estimate.

Default: normal kernel

Restriction: You can specify up to five values to request multiple estimates.

Requirement: You must enclose this suboption in parentheses after the KERNEL option.

Interaction: You can also use the K= suboption with the C= suboption, which specifies standardized bandwidths. If you specify more kernel functions than

bandwidths, PROC UNIVARIATE repeats the last bandwidth in the list for the remaining estimates. Likewise, if you specify more bandwidths than kernel functions, PROC UNIVARIATE repeats the last kernel function for the remaining estimates. For example, the following statements compute three estimates with bandwidths of 0.5, 1.0, and 1.5:

```
proc univariate;
   var length;
   histogram length / kernel(c=0.5 1.0 1.5 k=normal quadratic);
run;
```

The first estimate uses a normal kernel, and the last two estimates use a quadratic kernel.

KERNEL<(*kernel-suboptions*)>
superimposes up to five kernel density estimates on the histogram. By default, PROC UNIVARIATE uses the AMISE method to compute kernel density estimates.

Tip: To request multiple kernel density estimates on the same histogram, specify a list of values for either the C= suboption or K= suboption.

Main discussion: "Kernel Density Estimates" on page 1410

See also: C= suboption on page 1341 and K= suboption on page 1345

L=*linetype*
specifies the line type for a fitted density curve or kernel density estimate curve.

Default: 1, which produces a solid line.

Requirement: You must enclose the L= suboption in parentheses after a density curve option or the KERNEL option.

Interaction: If you use the L= suboption with the KERNEL option, you can specify a single line type or a list of line types.

See also: For a list of available line types, see *SAS/GRAPH Software: Reference*

Featured in: Example 8 on page 1444

LGRID=*linetype*
specifies the line type for the grid when a grid displays on the histogram.

Default: 1, which produces a solid line

Interaction: This option automatically invokes the GRID= option.

LHREF=*linetype*
specifies the line type for the reference lines that you request with the HREF= option.

Alias: LH=

Default: 2, which produces a dashed line

LOGNORMAL<(*lognormal-suboptions*)>
displays a fitted lognormal density curve on the histogram.

Restriction: The LOGNORMAL option can occur only once in a HISTOGRAM statement.

Interaction: The parameter θ must be less than the minimum data value. Use the THETA= suboption to specify θ. The default value for θ is zero. Specify THETA=EST to request the maximum likelihood estimate for θ.

Interaction: Use the SIGMA= and ZETA= suboptions to specify σ and ζ. By default, PROC UNIVARIATE computes a maximum likelihood estimate for σ and ζ. For example, the following statements fit a lognormal distribution function with a default value of $\theta = 0$ and with maximum likelihood estimates for σ and ζ:

```
proc univariate;
   var length;
```

```
   histogram length/ lognormal;
run;
```

Main discussion: See "Lognormal Distribution" on page 1408

See also: the ZETA= suboption on page 1352, SIGMA= suboption on page 1350, and THETA= suboption on page 1350

LVREF=*linetype*
specifies the line type for the reference lines that you request with the VREF= option.

Alias: LV=

Default: 2, which produces a dashed line

MAXNBIN=*n*
specifies the maximum number of bins in the comparative histogram that display. This option is useful when the scales or ranges of the data distributions differ greatly from cell to cell.

By default, PROC UNIVARIATE determines the bin size and midpoints for the key cell, and then extends the midpoint list to accommodate the data ranges for the remaining cells. However, if the cell scales differ considerably, the resulting number of bins may be so great that each cell histogram is scaled into a narrow region. By using MAXNBIN= to limit the number of bins, you can narrow the window about the data distribution in the key cell.

Requirement: This option is ignored unless you specify the CLASS statement.

Tip: MAXNBIN= provides an alternative to the MAXSIGMAS= option.

MAXSIGMAS=*value*
specifies to limit the number of bins in the comparative histogram that display to a range of *value* standard deviations (of the data in the key cell) above and below the mean of the data in the key cell. This option is useful when the scales or ranges of the data distributions differ greatly from cell to cell.

By default, PROC UNIVARIATE determines the bin size and midpoints for the key cell, and then extends the midpoint list to accommodate the data ranges for the remaining cells. However, if the cell scales differ considerably, the resulting number of bins may be so great that each cell histogram is scaled into a narrow region. By using MAXSIGMAS= to limit the number of bins, you can narrow the window that surrounds the data distribution in the key cell.

Requirement: This option is ignored unless you specify the CLASS statement.

MIDPERCENTS
requests a table that lists the midpoints and percentage of observations in each histogram interval.

Interaction: If you specify MIDPERCENTS in parentheses after a density estimate option, PROC UNIVARIATE displays a table that lists the midpoints, the observed percentage of observations, and the estimated percentage of the population in each interval (estimated from the fitted distribution).

MIDPOINTS=*value(s)*|KEY|UNIFORM
specifies how to determine the midpoints for the histogram intervals, where

value(s)
determines the width of the histogram bars as the difference between consecutive midpoints. PROC UNIVARIATE uses the same *value(s)* for all variables.

Range: The range of midpoints, extended at each end by half of the bar width, must cover the range of the data. For example, if you specify

```
midpoints=2 to 10 by 0.5
```

then all of the observations should fall between 1.75 and 10.25.

Requirement: You must use evenly spaced midpoints which you list in increasing order.

KEY
determines the midpoints for the data in the key cell. The initial number of midpoints is based on the number of observations in the key cell that use the method of Terrell and Scott (1985). PROC UNIVARIATE extends the midpoint list for the key cell in either direction as necessary until it spans the data in the remaining cells.

Requirement: This option is ignored unless you specify the CLASS statement.

UNIFORM
determines the midpoints by using all the observations as if there were no cells. In other words, the number of midpoints is based on the total sample size by using the method of Terrell and Scott (1985).

Requirement: This option does not apply unless you specify the CLASS statement.

Default: If you use a CLASS statement, MIDPOINTS=KEY; however, if the key cell is empty then MIDPOINTS=UNIFORM. Otherwise, PROC UNIVARIATE computes the midpoints by using an algorithm (Terrell and Scott, 1985) that is primarily applicable to continuous data that are approximately normally distributed.

Featured in: Example 8 on page 1444 and Example 10 on page 1450

MU=*value*
specifies the parameter μ for normal density curves.

Default: the sample mean

Requirement: You must enclose this suboption in parentheses after the NORMAL option.

NAME='*string*'
specifies a name for the plot, up to eight characters long, that appears in the PROC GREPLAY master menu.

Default: UNIVAR

NCOLS=*n*
specifies the number of columns in the comparative histogram.

Alias: NCOL=

Default: NCOLS=1, if you specify only one class variable, and NCOLS=2, if you specify two class variables.

Requirement: This option is ignored unless you specify the CLASS statement.

Interaction: If you specify two class variables, you can use the NCOLS= option with the NROWS= option.

Featured in: Example 10 on page 1450

NOBARS
suppresses drawing of histogram bars.

Tip: Use this option to display only the fitted curves.

NOFRAME
suppresses the frame that surrounds the subplot area.

NOHLABEL
suppresses the label for the horizontal axis.

Tip: Use this option to reduce clutter.

NOPLOT
suppresses the creation of a plot.

Alias: NOCHART

Tip: Use NOPLOT when you want to display only descriptive statistics for a fitted density or create an OUTHISTOGRAM= data set.

NOPRINT
suppresses the table of statistics that summarizes the fitted density curve.

Requirement: Enclose this option in the parentheses that follow the density curve option.

Featured in: Example 8 on page 1444

NORMAL<(*normal-suboptions*)>
displays a fitted lognormal density curve on the histogram.

Restriction: The NORMAL option can occur only once in a HISTOGRAM statement.

Interaction: Use the MU= and SIGMA= suboptions to specify μ and σ. By default, PROC UNIVARIATE uses the sample mean and sample standard deviation for μ and σ.

Main discussion: See "Normal Distribution" on page 1409

See also: the MU= suboption on page 1348 and the SIGMA= suboption on page 1350

Featured in: Example 8 on page 1444

NOVLABEL
suppresses the label for the vertical axis.

NOVTICK
suppresses the tick marks and tick mark labels for the vertical axis.

Interaction: This option automatically invokes the NOVLABEL option.

NROWS=*n*
specifies the number of rows in the comparative histogram.

Alias: NROW=

Default: 2

Requirement: This option is ignored unless you specify the CLASS statement.

Interaction: If you specify two class variables, you can use the NCOLS= option with the NROWS= option.

Featured in: Example 10 on page 1450

OUTHISTOGRAM=*SAS-data-set*
creates a SAS data set that contains information about histogram intervals. Specifically, the data set contains the midpoints of the histogram intervals, the observed percentage of observations in each interval, and the estimated percentage of observations in each interval (estimated from each of the specified fitted curves).

Alias: OUTHIST=

See also: "OUTHISTOGRAM= Data Set" on page 1417

PERCENTS=*value(s)*
specifies a list of percentages that PROC UNIVARIATE uses to calculate quantiles from the data and to estimate quantiles from the fitted density curve.

Alias: PERCENT=

Default: 1, 5, 10, 25, 50, 75, 90, 95, and 99 percent

Range: between 0 and 100

Requirement: You must enclose this suboption in parentheses after the curve option.

PFILL=*pattern*
specifies a pattern to fill the bars of the histograms (or the areas that are under a fitted density curve if you also specify the FILL option).

Default: The bars and curve areas are not filled.

See also: CFILL= option on page 1342 and FILL option on page 1343

See also: *SAS/GRAPH Software: Reference*

RTINCLUDE
includes the right endpoint of each histogram interval in that interval. By default, PROC UNIVARIATE includes the left endpoint in the histogram interval.

SCALE=*value*
is an alias for the SIGMA= suboption when you request density curves with the BETA, EXPONENTIAL, GAMMA, and WEIBULL options and an alias for the ZETA= suboption when you request density curves with the LOGNORMAL option.

See also: SIGMA= suboption on page 1350 and ZETA= suboption on page 1352

SHAPE=*value*
is an alias for the ALPHA= suboption when you request gamma curves with the GAMMA option, the SIGMA= suboption when you request lognormal curves with the LOGNORMAL option, and the C= suboption when you request Weibull curves with the WEIBULL option.

See also: ALPHA suboption on page 1340, SIGMA suboption on page 1350, and C= suboption on page 1341

SIGMA=*value*|EST
specifies the parameter σ for the fitted density curve when you request the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, and WEIBULL options. See Table 41.2 on page 1350 for a summary of how to use the SIGMA= suboption.

Default: see Table 41.2 on page 1350

Requirement: You must enclose this suboption in parentheses after the density curve option.

Tip: As a BETA suboption, you can specify SIGMA=EST to request a maximum likelihood estimate for σ.

Table 41.2 Uses of the SIGMA suboption

Distribution Keyword	SIGMA= Specifies	Default Value	Alias
BETA	scale parameter σ	1	SCALE=
EXPONENTIAL	scale parameter σ	maximum likelihood estimate	SCALE=
GAMMA	scale parameter σ	maximum likelihood estimate	SCALE=
WEIBULL	scale parameter σ	maximum likelihood estimate	SCALE=
LOGNORMAL	shape parameter σ	maximum likelihood estimate	SCALE=
NORMAL	scale parameter σ	standard deviation	SHAPE=

THETA=*value*|EST
specifies the lower threshold parameter θ for the fitted density curve when you request the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, and WEIBULL options.

Default: 0

Requirement: You must enclose this suboption in parentheses after the curve option.

Tip: To compute a maximum likelihood estimate for θ, specify THETA=EST.

THRESHOLD= ***value***
is an alias for the THETA= option. See the THETA= suboption on page 1350.

TURNVLABELS
specifies that PROC UNIVARIATE turn the characters in the vertical axis labels so that they display vertically. This happens by default when you use a hardware font.

Alias: TURNVLABEL

VAXIS=*value(s)*
specifies tick mark values for the vertical axis.

Requirement: Use evenly spaced values which you list in increasing order. The first value must be zero and the last value must be greater than or equal to the height of the largest bar. You must scale the values in the same units as the bars.

See also: the VSCALE= option on page 1351

Featured in: Example 10 on page 1450

VAXISLABEL='*label*'
specifies a label for the vertical axis.

Requirement: Labels can have up to 40 characters.

Featured in: Example 10 on page 1450

VMINOR=*n*
specifies the number of minor tick marks between each major tick mark on the vertical axis. PROC UNIVARIATE does not label minor tick marks.

Alias: VM=

Default: 0

VOFFSET=*value*
specifies the offset in percentage screen units at the upper end of the vertical axis.

VREF=*value(s)*
draws reference lines that are perpendicular to the vertical axis at the *value(s)* that you specify.

See also: CVREF= option on page 1343 and LVREF= option on page 1347.

VREFLABELS=' *label1*'... '*labeln*'
specifies labels for the reference lines that you request with the VREF= option.

Alias: VREFLABEL= and VREFLAB=

Restriction: The number of labels must equal the number of reference lines. Labels can have up to 16 characters.

VREFLABPOS=*n*
specifies the horizontal position of VREFLABELS= labels, where *n* is

1	positions the labels at the left of the histogram.
2	positions the labels at the right of the histogram.

Default: 1

VSCALE=*scale*
specifies the scale of the vertical axis, where *scale* is

COUNT
scales the data in units of the number of observations per data unit.

PERCENT
scales the data in units of percentage of observations per data unit.

PROPORTION
scales the data in units of proportion of observations per data unit.

Default: PERCENT

Featured in: Example 10 on page 1450

W=*n*
specifies the width in pixels of the fitted density curve or the kernel density estimate curve.

Default: 1

Requirement: You must enclose this suboption in parentheses after the density curve option or the KERNEL option.

Interaction: As a KERNEL suboption, you can specify a list of up to five W= values.

WAXIS=*n*
specifies the line thickness (in pixels) for the axes and frame.

Default: 1

WBARLINE=*n*
specifies the line thickness for the histogram bar outlines.

Default: 1

WEIBULL<(*Weibull-suboptions*)>
displays a fitted Weibull density curve on the histogram.

Restriction: The WEIBULL option can occur only once in a HISTOGRAM statement.

Interaction: The parameter θ must be less than the minimum data value. Use the THETA= suboption to specify θ. The default value for θ is zero. Specify THETA=EST to request the maximum likelihood estimate for θ.

Interaction: Use ALPHA= and the SIGMA= suboptions to specify the shape parameter c and the scale parameter σ. By default, PROC UNIVARIATE computes the maximum likelihood estimates for c and σ. For example, the following statements fit a Weibull curve with $\theta = 15$ and with a maximum likelihood estimate for c and σ:

```
proc univariate;
   var length;
   histogram length/ weibull(theta=4);
run;
```

PROC UNIVARIATE calculates the maximum likelihood estimate of α iteratively by using the Newton-Raphson approximation.

Main discussion: See "Weibull Distribution" on page 1409

See also: the C= suboption on page 1341, SIGMA= suboption on page 1350, and THETA= suboption on page 1350

WGRID=*n*
specifies the line thickness for the grid.

ZETA= *value*
specifies a value for the scale parameter ζ for the lognormal density curve when you request the LOGNORMAL option.

Default: a maximum likelihood estimate

Requirement: You must enclose this suboption in parentheses after the LOGNORMAL option.

ID Statement

Identifies the extreme observations in the table of extreme observations.

Featured in: Example 2 on page 1421

ID *variable(s)*;

Required Arguments

variable(s)
specifies one or more variables to include in the table of extreme observations. The corresponding values of the ID variables appear beside the n largest and n smallest observations, where n is the value of NEXTROBS= option.

See also: NEXTROBS= on page 1329

INSET Statement

Places a box or table of summary statistics, called an *inset*, directly in the high-resolution graph.

Requirement: The INSET statement must follow the HISTOGRAM, PROBPLOT, or QQPLOT statement that creates the plot that you want to augment. The inset appears in all the graphs that the preceding plot statement produces.

Tip: You can use multiple INSET statements.

Featured in: Example 9 on page 1448 and Example 10 on page 1450

INSET <*keyword(s)* DATA=*SAS-data-set*> </ *option(s)*>;

Arguments

keyword(s)
specifies one or more keywords that identify the information to display in the inset. PROC UNIVARIATE displays the information in the order that you request the keywords.

You can specify statistical keywords, primary keywords, and secondary keywords. The available statistical keywords are

Descriptive statistic keywords		
CSS	CV	KURTOSIS
MAX	MEAN	N
MIN	MODE	RANGE
NMISS	NOBS	STDMEAN
SKEWNESS	STD	USS
SUM	SUMWGT	VAR
Quantile statistic keywords		
MEDIAN	P1	P5
P10	P90	P95
P99	Q1	Q3
QRANGE		
Robust statistic keywords		
GINI	MAD	QN
SN	STD_GINI	STD_MAD
STD_QN	STD_QRANGE	STD_SN
Hypothesis testing keywords		
MSIGN	PROBM	PROBT
NORMALTEST	PROBN	SIGNRANK
PNORMAL	PROBS	T

A *primary keyword* allows you to specify *secondary keywords* in parentheses immediately after the primary keyword. Primary keywords are BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, WEIBULL, WEIBULL2, KERNEL, and KERNEL*n*. If you specify a primary keyword but omit a secondary keyword, the inset displays a colored line and the distribution name as a key for the density curve. For a list of the secondary keywords, see Table 41.3 on page 1354.

By default, PROC UNIVARIATE identifies inset statistics with appropriate labels and prints numeric values using appropriate formats. To customize the label, specify the keyword followed by an equal sign (=) and the desired label in quotes. To customize the format, specify a numeric format in parentheses after the keyword. Labels can have up to 24 characters. If you specify both a label and a format for a statistic, the label must appear before the format. For example,

```
inset n='Sample Size' std='Std Dev' (5.2);
```

requests customized labels for two statistics and displays the standard deviation with field width of 5 and two decimal places.

Table 41.3 Available Secondary Keywords

Keyword	Alias	Description
For BETA primary keyword		
ALPHA	SHAPE1	first shape parameter α

Keyword	Alias	Description
BETA	SHAPE2	second shape parameter β
SIGMA	SCALE	scale parameter σ
THETA	THRESHOLD	lower threshold parameter θ
For EXP primary keyword		
SIGMA	SCALE	scale parameter σ
THETA	THRESHOLD	threshold parameter θ
For GAMMA primary keyword		
ALPHA	SHAPE	shape parameter α
SIGMA	SCALE	scale parameter σ
THETA	THRESHOLD	threshold parameter θ
For LOGNORMAL primary keyword		
SIGMA	SHAPE	shape parameter σ
THETA	THRESHOLD	threshold parameter θ
ZETA	SCALE	scale parameter ζ
For NORMAL primary keyword		
MU	MEAN	mean parameter μ
SIGMA	STD	shape parameter σ
For WEIBULL primary keyword		
C	SHAPE	shape parameter c
SIGMA	SCALE	scale parameter σ
THETA	THRESHOLD	threshold parameter θ
For WEIBULL2 primary keyword		
C	SHAPE	shape parameter c
SIGMA	SCALE	scale parameter σ
THETA	THRESHOLD	known lower threshold parameter θ_0
For any parametric distribution primary keyword*		
AD		Anderson-Darling EDF test statistic
ADPVAL		Anderson-Darling EDF test *p*-value
CVM		Cramer-von Mises EDF test statistic
CVMPVAL		Cramer-von Mises EDF test *p*-value
KSD		Kolmogorov-Smirnov EDF test statistic
KSDPVAL		Kolmogorov-Smirnov EDF test *p*-value
For KERNEL or KERNEL*n* primary keyword*		
TYPE		kernel type: normal, quadratic, or triangular

Keyword	Alias	Description
BANDWIDTH	BWIDTH	bandwidth λ for the density estimate
C		standardized bandwidth c for the density estimate: $c = \frac{\lambda}{Q} n^{\frac{1}{5}}$ where n =sample size, λ =bandwidth, and Q =interquartile range
AMISE		approximate mean integrated square error (MISE) for the kernel density

* Available with only the HISTOGRAM statement and a BETA, EXPONENTIAL, LOGNORMAL, NORMAL, or WEIBULL distribution.

Requirement: Some inset statistics are not available unless you request a plot statement and options that calculate these statistics. For example:

```
proc univariate data=score;
    histogram final / normal;
    inset mean std normal(ad adpval);
run;
```

The MEAN and STD keywords display the sample mean and standard deviation of FINAL. The NORMAL keyword with the *secondary keywords* AD and ADPVAL display the Anderson-Darling goodness-of-fit test statistic and *p*-value. The statistics that are specified with the NORMAL keyword are available only because the NORMAL option is requested in the HISTOGRAM statement.

The KERNEL or KERNEL*n* keyword is available only if you request a kernel density estimate in a HISTOGRAM statement. The WEIBULL2 keyword is available only if you request a two-parameter Weibull distribution in the PROBPLOT or QQPLOT statement.

Tip: To specify the same format for all the statistics in the INSET statement, use the FORMAT= option.

Tip: To create a completely customized inset, use a DATA= data set. The data set contains the label and the value that you want to display in the inset.

Tip: If you specify multiple kernel density estimates, you can request inset statistics for all the estimates with the KERNEL keyword. Alternatively, you can display inset statistics for individual curves with KERNEL*n* keyword, where *n* is the curve number between 1 and 5.

Featured in: Example 9 on page 1448 and Example 10 on page 1450

DATA=*SAS-data-set*

requests that PROC UNIVARIATE display customized statistics from a SAS data set in the inset table. The data set must contain two variables:

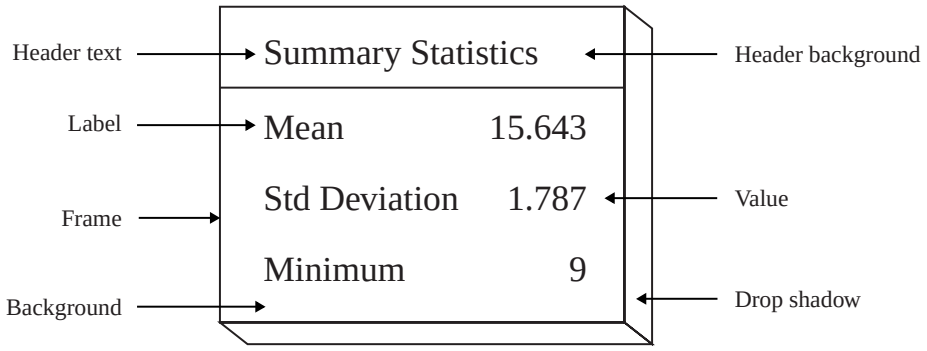
LABEL	a character variable whose values provide labels for inset entries.
VALUE	a variable that is either character or numeric and whose values provide values for inset entries.

The label and value from each observation in the data set occupy one line in the inset. The position of the DATA= keyword in the keyword list determines the position of its lines in the inset.

Options

Figure 41.1 on page 1357 illustrates the meaning of terms that are used in this section.

Figure 41.1 The Inset



CFILL=*color* | BLANK
specifies the color of the background which, if you omit the CFILLH= option, includes the header background.

Default The background is empty which causes items that overlap the inset (such as curves, histogram bars, or specification limits) to show through the inset.

Tip: Specify a value for CFILL= so that items that overlap no longer show through the inset. Use CFILL=BLANK to leave the background uncolored.

CFILLH=*color*
specifies the color of the header background.

Default: the CFILL= *color*

CFRAME=*color*
specifies the color of the frame.

Default: the same color as the axis of the plot

CHEADER=*color*
specifies the color of the header text.

Default: the CTEXT=*color*

CSHADOW=*color*
specifies the color of the drop shadow.

Default: A drop shadow is not displayed.

CTEXT=*color*
specifies the color of the text.

Default: the same color as the other text on the plot

DATA
specifies how to use data coordinates to position the inset with the POSITION= option.

Requirement: The DATA option is available only when you specify POSITION=(*x*,*y*). You must place DATA immediately after the coordinates (*x*,*y*).

Main Discussion: "Positioning the Inset Using Coordinates" on page 1360

See also: POSITION= option on page 1358

FONT=*font*
specifies the font of the text.

Default: If you locate the inset in the interior of the plot then the font is SIMPLEX. If you locate the inset in the exterior of the plot then the font is the same as the other text on the plot.

Featured in: Example 10 on page 1450

FORMAT=*format*
specifies a format for all the values in the inset.

Interaction: If you specify a format for a particular statistic, then this format overrides FORMAT=*format*.

See also: For more information about SAS formats, see *SAS Language Reference: Dictionary*

Featured in: Example 9 on page 1448

HEADER=*string*
specifies the header text where *string* cannot exceed 40 characters.

Default: No header line appears in the inset.

Interaction: If all the keywords that you list in the INSET statement are secondary keywords that correspond to a fitted curve on a histogram, PROC UNIVARIATE displays a default header that indicates the distribution and identifies the curve.

Featured in: Example 9 on page 1448

HEIGHT=*value*
specifies the height of the text.

Featured in: Example 10 on page 1450

NOFRAME
suppresses the frame drawn around the text.

Featured in: Example 10 on page 1450

POSITION=*position*
determines the position of the inset. The *position* is a compass point keyword, a margin keyword, or a pair of coordinates (*x*,*y*).

Alias: POS=

Default: NW, which positions the inset in the upper left (northwest) corner of the display.

Requirement: You must specify coordinates in axis percentage units or axis data units.

Main discussion: "Positioning the Inset Using Compass Point" on page 1359, "Positioning the Inset in the Margins" on page 1359, and "Positioning the Inset Using Coordinates" on page 1360

Featured in: Example 9 on page 1448 and Example 10 on page 1450

REFPOINT=BR | BL | TR | TL
specifies the reference point for an inset that PROC UNIVARIATE positions by a pair of coordinates with the POSITION= option. The REFPOINT= option specifies which corner of the inset frame that you want to position at coordinates (*x*,*y*). The reference points are

BL	bottom left
BR	bottom right
TL	top left
TR	top right

Default: BL

Requirement: You must use REFPOINT= with POSITION=(*x*,*y*) coordinates.

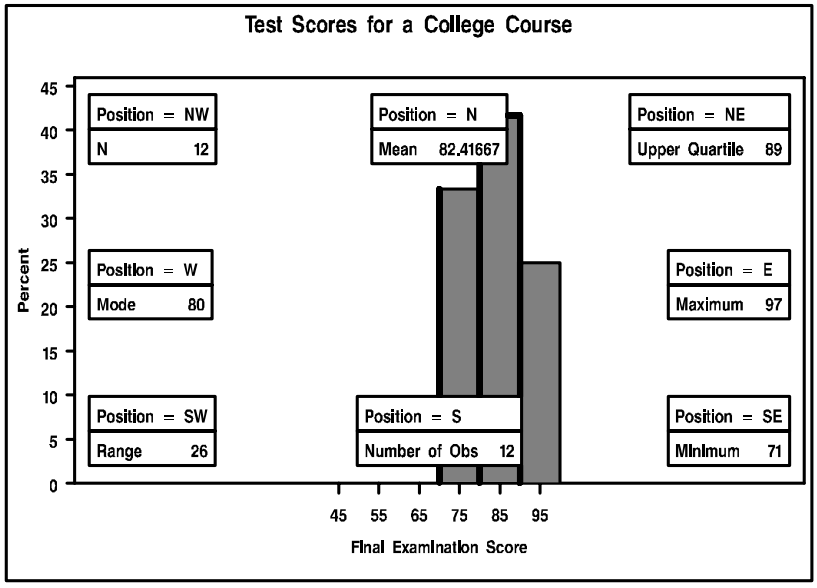
Featured in: Example 9 on page 1448

Positioning the Inset Using Compass Point

To position the inset by using a compass point position, use the keyword N, NE, E, SE, S, SW, W, or NW in the POSITION= option. The default position of the inset is NW.

The following statements produce a histogram to show the position of the inset for the eight compass points:

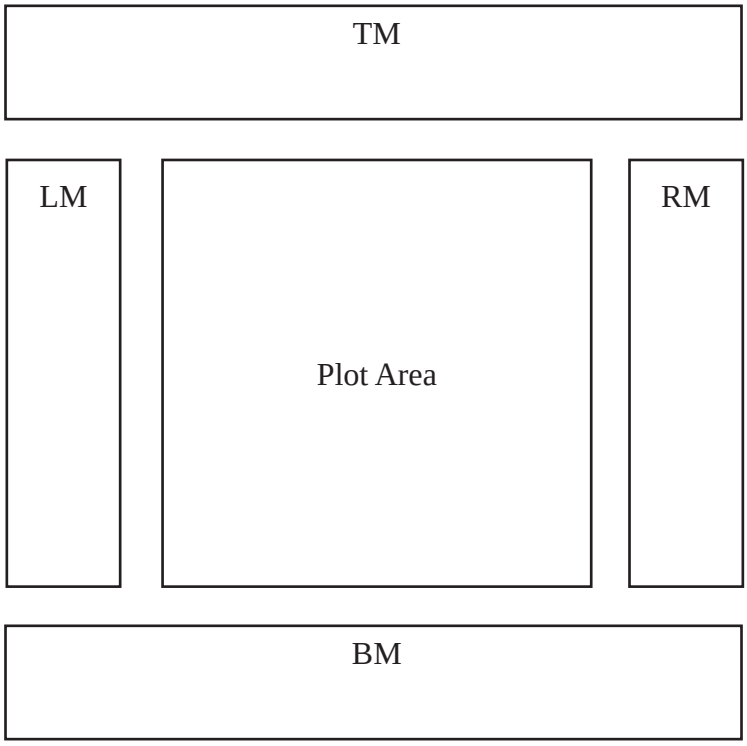
```
proc univariate data=score noprint;
   histogram final / cfill=gray midpoints=45 to 95 by 10 barwidth=5;
   inset n     / cfill=blank header='Position = NW' pos=nw;
   inset mean  / cfill=blank header='Position = N ' pos=n ;
   inset sum   / cfill=blank header='Position = NE' pos=ne;
   inset max   / cfill=blank header='Position = E ' pos=e ;
   inset min   / cfill=blank header='Position = SE' pos=se;
   inset nobs  / cfill=blank header='Position = S ' pos=s ;
   inset range / cfill=blank header='Position = SW' pos=sw;
   inset mode  / cfill=blank header='Position = W ' pos=w ;
   label final='Final Examination Score';
   title 'Test Scores for a College Course';
run;
```



Positioning the Inset in the Margins

To position the inset in one of the four margins that surround the plot area use the margin keywords LM, RM, TM, or BM in the POSITION= option. Figure 41.2 on page 1360 shows the location of the inset in the margin.

Figure 41.2 Locating the Inset in the Margins



Margin positions are recommended if you list a large number of statistics in the INSET statement. If you attempt to display a lengthy inset in the interior of the plot, it is most likely that the inset will collide with the data display.

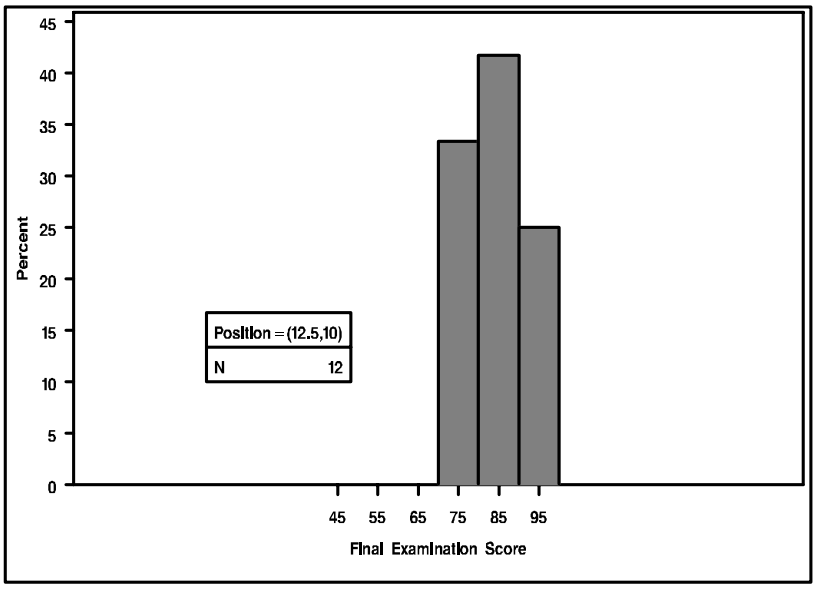
Positioning the Inset Using Coordinates

To position the inset with coordinates, use POSITION=(*x,y*). You specify the coordinates in axis data units or in axis percentage units (the default).

data unit

If you specify the DATA option immediately following the coordinates, PROC UNIVARIATE positions the inset by using axis data units. For example, the following statements place the bottom left corner of the inset at 12.5 on the horizontal axis and 10 on the vertical axis:

```
proc univariate data=score;
   histogram final / midpoints 45 to 95 by 10 barwidth=5
                     cfill=gray ;
   inset n / header   = 'Position=(12.5,10)'
             position = (12.5,10) data;
run;
```



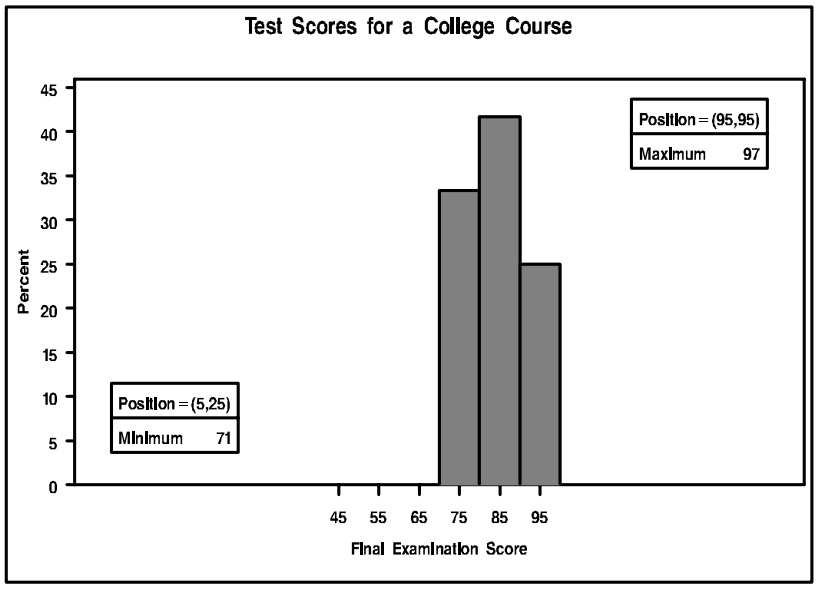
By default, the specified coordinates determine the position of the bottom left corner of the inset. To change this reference point, use the REFPOINT= option (see the next example).

axis percent unit

If you omit the DATA option, PROC UNIVARIATE positions the inset by using axis percentage units. The coordinates in axis percentage units must be *between* 0 and 100. The coordinates of the bottom left corner of the display are (0,0), while the upper right corner is (100,100). For example, the following statements create a histogram and use coordinates in axis percentage units to position the two insets:

```
proc univariate data=sccore;
   histogram final / midpoints 45 to 95 by 10 barwidth=5
                     cfill=gray;
   inset min / position = (5,25)
               header   = 'Position=(5,25)'
               refpoint = tl;
   inset max / position = (95,95)
               header   = 'Position=(95,95)'
               refpoint = tr;
   run;
```

The REFPOINT= option determines which corner of the inset to place at the coordinates that are specified with the POSITION= option. The first inset uses REFPOINT=TL, so that the top left corner of the inset is positioned 5% of the way across the horizontal axis and 25% of the way up the vertical axis. The second inset uses REFPOINT=TR, so that the top right corner of the inset is positioned 95% of the way across the horizontal axis and 95% of the way up the vertical axis.



OUTPUT Statement

Saves statistics and BY variables in an output data set.

Tip: You can save percentiles that are not automatically computed.

Tip: You can use multiple OUTPUT statements to create several OUT= data sets.

Main discussion: "Output Data Set" on page 1417

Featured in: Example 5 on page 1432, Example 6 on page 1437, and Example 7 on page 1439

OUTPUT <OUT=*SAS-data-set*> *statistic-keyword-1=name(s)*
<*…statistic-keyword-n=name(s)*> <*percentiles-specification*> ;

Options

OUT=*SAS-data-set*
identifies the output data set. If *SAS-data-set* does not exist, PROC UNIVARIATE creates it. If you omit OUT=, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique.

Default: DATA*n*

statistic-keyword=name(s)
specifies a statistic to store in the OUT= data set and names the new variable that will contain the statistic. The available statistical keywords are

Descriptive statistic keywords

CSS	CV	KURTOSIS
MAX	MEAN	N
MIN	MODE	RANGE

NMISS	NOBS	STDMEAN
SKEWNESS	STD	USS
SUM	SUMWGT	VAR

Quantile statistic keywords

MEDIAN	P1	P5
P10	P90	P95
P99	Q1	Q3
QRANGE		

Robust statistic keywords

GINI	MAD	QN
SN	STD_GINI	STD_MAD
STD_QN	STD_QRANGE	STD_SN

Hypothesis testing keywords

NORMAL	PROBN	MSIGN
PROBM	SIGNRANK	PROBS
T	PROBT	

See Appendix 1, "SAS Elementary Statistics Procedures," on page 1457 and "Statistical Computations" on page 1393 for the keyword definitions and statistical formulas.

To store the same statistic for several analysis variables, specify a list of names. The order of the names corresponds to the order of the analysis variables in the VAR statement. PROC UNIVARIATE uses the first name to create a variable that contains the statistic for the first analysis variable, the next name to create a variable that contains the statistic for the second analysis variable, and so on. If you do not want to output statistics for all the analysis variables, specify fewer names than the number of analysis variables.

percentiles-specification

specifies one or more percentiles to store in the OUT= data set and names the new variables that contain the percentiles. The form of *percentiles-specification* is

PCTLPTS=*percentile(s)* PCTLPRE=*prefix-name(s)* <PCTLNAME=*suffix-name(s)*>

PCTLPTS=*percentile(s)*

specifies one or more percentiles to compute. You can specify percentiles with the expression *start* TO *stop* BY *increment* where *start* is a starting number, *stop* is an ending number, and *increment* is a number to increment by.

Range: any decimal numbers between 0 and 100, inclusive

Example: To compute the 50th, 95th, 97.5th, and 100th percentiles, submit the statement

```
output pctlpre=P_ pctlpts=50,95 to 100 by 2.5;
```

PCTLPRE=*prefix-name(s)*

specifies one or more prefixes to create the variable names for the variables that contain the PCTLPTS= percentiles. To save the same percentiles for more than one analysis variable, specify a list of prefixes. The order of the prefixes corresponds to the order of the analysis variables in the VAR statement.

Interaction: PROC UNIVARIATE creates a variable name by combining the PCTLPRE= value and either *suffix-name* or (if you omit PCTLNAME= or if you specify too few *suffix-name(s)*) the PCTLPTS= value.

PCTLNAME=*suffix-name(s)*
specifies one or more suffixes to create the names for the variables that contain the PCTLPTS= percentiles. PROC UNIVARIATE creates a variable name by combining the PCTLPRE= value and *suffix-name*. Because the suffix names are associated with the percentiles that are requested, list the suffix names in the same order as the PCTLPTS= percentiles.

Requirement: You must specify PCTLPRE= to supply prefix names for the variables that contain the PCTLPTS= percentiles.

Interaction: If the number of PCTLNAME= values is fewer than the number of *percentile(s)* or if you omit PCTLNAME=, PROC UNIVARIATE uses *percentile* as the suffix to create the name of the variable that contains the percentile. For an integer percentile, PROC UNIVARIATE uses *percentile*. For a noninteger percentile, PROC UNIVARIATE truncates decimal values of *percentile* to two decimal places and replaces the decimal point with an underscore.

Interaction: If either the prefix and suffix name combination or the prefix and percentile name combination is longer than 32 characters, PROC UNIVARIATE truncates the prefix name so that the variable name is 32 characters.

Saving Percentiles Not Automatically Computed

You can use PCTLPTS= to output percentiles that are not in the list of quantile statistics. PROC UNIVARIATE computes the requested percentiles based on the method that you specify with the PCTLDEF= option in the PROC UNIVARIATE statement. You must use PCTLPRE=, and optionally PCTLNAME=, to specify variable names for the percentiles. For example, the following statements create an output data set that is named PCTLS that contains the 20th and 40th percentiles of the analysis variables Test1 and Test2:

```
proc univariate data=score;
   var Test1 Test2;
   output out=pctls pctlpts=20 40 pctlpre=Test1_ Test2_
              pctlname=P20 P40;
run;
```

PROC UNIVARIATE saves the 20th and 40th percentiles for Test1 and Test2 in the variables Test1_P20, Test2_P20, Test1_P40, and Test2_P40.

Using the BY Statement with the OUTPUT Statement

When you use a BY statement, the number of observations in the OUT= data set corresponds to the number of BY groups. Otherwise, the OUT= data set contains only one observation.

PROBPLOT Statement

Creates a probability plot by using high-resolution graphs, which compare ordered variable values with the percentiles of a specified theoretical distribution.

Alias: PROB

Default: Normal probability plot

Restriction: You can specify only one theoretical distribution.

Tip: You can use multiple PROBPLOT statements.

Main discussion:

Featured in: "Quantile–Quantile and Probability Plots" on page 1391

PROBPLOT <*variable(s)*> </ *option(s)*>;

To do this:	Use this option:
Request a distribution	
Specify beta probability plot with required shape parameters α, β.	BETA(*beta-suboptions)*
Specify exponential probability plot	EXPONENTIAL(*exponential-suboptions)*
Specify gamma probability plot with a required shape parameter α	GAMMA(*gamma-suboptions*)
Specify lognormal probability plot with a required shape parameter σ	LOGNORMAL(*lognormal-suboptions*)
Specify normal probability plot	NORMAL(*normal-suboptions)*
Specify three-parameter Weibull probability plot with a required shape parameter c	WEIBULL(*Weibull-suboptions*)
Specify two-parameter Weibull probability plot	WEIBULL2(*Weibull2-suboptions*)
Distribution suboptions	
Specify shape parameter α for the beta or gamma distribution	ALPHA=
Specify shape parameter β for the beta distribution	BETA=
Specify shape parameter c for the Weibull distribution or c_0 for distribution reference line of the Weibull2 distribution	C=
Specify μ_0 for distribution reference line for the normal distribution	MU=
Specify σ_0 for distribution reference line for the beta, exponential, gamma, normal, Weibull, or Weibull2 distribution or the required shape parameter σ for the lognormal option	SIGMA=
Specify slope of distribution reference line for the lognormal or Weibull2 distribution	SLOPE=

To do this:	Use this option:
Specify θ_0 for distribution reference line for the beta, exponential, gamma, lognormal, or Weibull distribution, or the lower known threshold θ_0 for the Weibull2 distribution	THETA=
Specify ζ_0 for distribution reference line for the lognormal distribution	ZETA=
Control appearance of distribution reference line	
Specify color of distribution reference line	COLOR=
Specify line type of distribution reference line	L=
Specify width of distribution reference line	W=
Control general plot layout	
Create a grid	GRID
Specify reference lines perpendicular to the horizontal axis	HREF=
Specify labels for HREF lines	HREFLABELS=
Specify a line style for grid lines	LGRID=
Adjust sample size when computing percentiles	NADJ=
Suppress frame around plotting area	NOFRAME
Request minor tick marks for percentile axis	PCTLMINOR
Specify tick mark labels for percentile axis	PCTLORDER=
Adjust ranks when computing percentiles	RANKADJ=
Display plot in square format	SQUARE
Specify reference lines perpendicular to the vertical axis	VREF=
Specify labels for VREF lines	VREFLABELS=
Enhance the probability plot	
Specify annotate data set	ANNOTATE=
Specify color for axis	CAXIS=
Specify color for frame	CFRAME=
Specify color for HREF= lines	CHREF=
Specify color for text	CTEXT=
Specify color for VREF= lines	CVREF=
Specify description for plot in graphics catalog	DESCRIPTION=
Specify software font for text	FONT=
Specify number of horizontal minor tick marks	HMINOR=
Specify line style for HREF= lines	LHREF=
Specify line style for VREF= lines	LVREF=
Specify name for plot in graphics catalog	NAME=
Specify number of vertical minor tick marks	VMINOR=

To do this:	Use this option:
Enhance the comparative probability plot	
Apply annotation requested in ANNOTATE= data set to key cell only	ANNOKEY
Specify color for filling frame for row labels	CFRAMESIDE=
Specify color for filling frame for column labels	CFRAMETOP=
Specify distance between tiles	INTERTILE=
Specify number of columns in comparative probability plot	NCOLS=
Specify number of rows in comparative probability plot	NROWS=

Arguments

variable(s)
identifies one or more variables that the procedure uses to create probability plots.

Default: If you omit *variable(s)* in the PROBPLOT statement then the procedure creates a probability plot for each variable that you list in the VAR statement, or for each numeric variable in the DATA= data set if you omit a VAR statement.

Requirement: If you specify a VAR statement, use a subset of the *variable(s)* that you list in the VAR statement. Otherwise, *variable(s)* are any numeric variables in the DATA= data set.

Options

ALPHA=*value(s)*|EST
specifies the required shape parameter $\alpha(\alpha > 0)$ for probability plots when you request the BETA or GAMMA options. The PROBPLOT statement creates a plot for each value that you specify.

Requirement: Enclose this suboption in parentheses following the BETA or GAMMA options.

Tip: To compute a maximum likelihood estimate for α, specify ALPHA=EST.

ANNOKEY
specifies to apply the annotation requested with the ANNOTATE= option to the *key cell* only. By default, PROC UNIVARIATE applies annotation to all of the cells.

Requirement: This option is ignored unless you specify the CLASS statement.

Tip: Use the KEYLEVEL= option in the CLASS statement to specify the key cell.

See also: the KEYLEVEL= option on page 1334

ANNOTATE=*SAS-data-set*
specifies an input data set that contains annotate variables as described in *SAS/GRAPH Software: Reference*.

Alias: ANNO=

Tip: The ANNOTATE = data set that you specify in the PROBPLOT statement is used by all plots that this statement creates. You can also specify an ANNOTATE= data set in the PROC UNIVARIATE statement to enhance all the graphics displays that the procedure creates.

See also: the ANNOTATE= option on page 1327 in the PROC UNIVARIATE statement

BETA(ALPHA=*value(s)*|EST BETA=*value(s)*|EST <*beta-suboptions*>)
displays a beta probability plot for each combination of the required shape parameters α and β.

Requirement: You must specify the shape parameters with the ALPHA= and BETA= suboptions.

Interaction: To create a plot that is based on maximum likelihood estimates for α and β, specify ALPHA=EST and BETA=EST.

Tip: To obtain graphical estimates of α and β, specify lists of values in the ALPHA= and BETA= suboptions. Then select the combination of α and β that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to the lower threshold parameter θ_0 and the scale parameter σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of θ_0 and σ_0 with THETA=EST and SIGMA=EST.

Agreement between the reference line and the point pattern indicates that the beta distribution with parameters α, β, θ_0, and σ_0 is a good fit.

Main discussion: "Beta Distribution" on page 1412

See also: the ALPHA= on page 1367 suboption and BETA= suboption on page 1368

BETA=*value(s)*|EST
specifies the shape parameter $\beta\,(\beta > 0)$ for probability plots when you request the BETA distribution option. PROC UNIVARIATE creates a plot for each value that you specify.

Alias: B=

Requirement: Enclose this suboption in parentheses after the BETA option.

Tip: To compute a maximum likelihood estimate for β, specify BETA=EST.

C=*value(s)*|EST
specifies the shape parameter $c\,(c > 0)$ for probability plots when you request the WEIBULL option or WEIBULL2 option. C= is a required suboption in the WEIBULL option.

Requirement: Enclose this suboption in parentheses after the WEIBULL option or WEIBULL2 option.

Interaction: To request a distribution reference line in the WEIBULL2 option, you must specify both the C= and SIGMA= suboptions.

Tip: To compute a maximum likelihood estimate for c, specify C=EST.

CAXIS=*color*
specifies the color for the axes.

Alias: CAXES=

Default: the first color in the device color list

Interaction: This option overrides any COLOR= specification.

CFRAME=*color*
specifies the color for the area that is enclosed by the axes and frame.

Default: the area is not filled

CFRAMESIDE=*color*
specifies the color to fill the frame area for the row labels that display along the left side of the comparative probability plot. This color also fills the frame area for the label of the corresponding class variable (if you associate a label with the variable).

Default: These areas are not filled.

Requirement: This option is ignored unless you specify the CLASS statement.

CFRAMETOP=*color*
specifies the color to fill the frame area for the column labels that display across the top of the comparative probability plot. This color also fills the frame area for the label of the corresponding class variable (if you associate a label with the variable).

Default: These areas are not filled.

Requirement: This option does not apply unless you specify the CLASS statement.

CHREF=*color*
specifies the color for horizontal axis reference lines when you specify the HREF= option.

Default: the first color in the device color list

COLOR=*color*
specifies the color of the diagonal distribution reference line.

Default: the first color in the device color list

Requirement: You must enclose this suboption in parentheses after a distribution option keyword.

CTEXT=*color*
specifies the color for tick mark values and axis labels.

Default: the color that you specify for the CTEXT= option in the GOPTIONS statement. If you omit the GOPTIONS statement, the default is the first color in the device color list.

CVREF=*color*
specifies the color for the reference lines that you request with the VREF= option.

Alias: CV=

Default: the first color in the device color list

DESCRIPTION='*string*'
specifies a description, up to 40 characters long, that appears in the PROC GREPLAY master menu.

Alias: DES=

Default: the variable name

EXPONENTIAL<(*exponential-options*)>
displays an exponential probability plot.

Alias: EXP

Tip: To assess the point pattern, add a diagonal distribution reference line that corresponds to θ_0 and σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of the threshold parameter θ_0 and the scale parameter σ_0 with the THETA=EST and SIGMA=EST suboptions.

Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters θ_0 and σ_0 is a good fit.

Main discussion: "Exponential Distribution" on page 1412

See also: the SIGMA= suboption on page 1373 and the THETA= suboption on page 1374

FONT=*font*
specifies a software font for the reference lines and the axis labels.

Default: hardware characters

Interaction: FONT=*font* takes precedence over the FTEXT=*font* that you specify in the GOPTIONS statement.

GAMMA(ALPHA=*value(s)*|EST <*gamma-suboptions*>)
displays a gamma probability plot for each value of the required shape parameter α.

Requirement: You must specify the shape parameter with the ALPHA= suboption.

Interaction: To create a plot that is based on a maximum likelihood estimate for α, specify ALPHA=EST.

Tip: To obtain a graphical estimate of α, specify a list of values in the ALPHA= suboption. Then select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to the threshold parameter θ_0 and the scale parameter σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of θ_0 and σ_0 with THETA=EST and SIGMA=EST.

Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters α, θ_0, and σ_0 is a good fit.

Main discussion: "Gamma Distribution" on page 1412

See also: the ALPHA= on page 1367 suboption, SIGMA suboption on page 1373, and THETA suboption on page 1374

GRID
displays a grid, drawing reference lines that are perpendicular to the percentile axis at major tick marks.

Default: 1

HMINOR=*n*
specifies the number of minor tick marks between each major tick mark on the horizontal axis. PROC UNIVARIATE does not label minor tick marks.

Alias: HM=

Default: 0

HREF=*value(s)*
draws reference lines that are perpendicular to the horizontal axis at the values you specify.

See also: CHREF= option on page 1369

HREFLABELS='*label1*' ... '*labeln*'
specifies labels for the reference lines that you request with the HREF= option.

Alias: HREFLABEL= and HREFLAB=

Restriction: The number of labels must equal the number of reference lines. Labels can have up to 16 characters.

HREFLABPOS=*n*
specifies the vertical position of HREFLABELS= labels, where *n* is

1	positions the labels at the left of the plot
2	positions the labels along the top of the plot
3	positions the labels from top to bottom

Default: 1

INTERTILE=*value*
specifies the distance in horizontal percentage screen units between the framed areas, which are called *tiles*.

Default: The tiles are contiguous.

Requirement: This option is ignored unless you specify the CLASS statement.

L=*linetype*
specifies the line type for a diagonal distribution reference line.

Default: 1, which produces a solid line

Requirement: You must enclose this suboption in parentheses after a distribution option.

LGRID=*linetype*
specifies the line type for the grid that you request with the GRID= option.

Default: 1, which produces solid lines

LHREF=*linetype*
specifies the line type for the reference lines that you request with the HREF= option.

Alias: LH=

Default: 2, which produces a dashed line

LOGNORMAL(SIGMA=*value(s)*|EST <*lognormal-suboptions*>)
displays a lognormal probability plot for each value of the required shape parameter σ.

Alias: LNORM

Requirement: You must specify the shape parameter with the SIGMA= suboption.

Interaction: To compute a maximum likelihood estimate for σ, specify SIGMA=EST.

Tip: To obtain a graphical estimate of σ, specify a list of values for the SIGMA= suboption, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to the threshold parameter θ_0 and the scale parameter ζ_0 with the THETA= and ZETA= suboptions. Alternatively, you can add a line that corresponds to estimated values of θ_0 and ζ_0 with THETA=EST and ZETA=EST.

Agreement between the reference line and the point pattern indicates that the lognormal distribution with parameters σ, θ_0, and ζ_0 is a good fit.

Main discussion: "Lognormal Distribution" on page 1413

See also: the SIGMA= suboption on page 1373, SLOPE= suboption on page 1373, THETA= suboption on page 1374, and ZETA= suboption on page 1375

LVREF=*linetype*
specifies the line type for the reference lines that you request with the VREF= option.

Default: 2, which produces a dashed line

MU=*value*|EST
specifies the mean μ_0 for a normal probability plot requested with the NORMAL option.

Default: the sample mean

Requirement: You must enclose this suboption in parentheses after the NORMAL option.

Tip: Specify the MU= and SIGMA= suboptions together to request a distribution reference line. Specify MU=EST to request a distribution reference line with μ_0 equal to the sample mean.

Featured in: Example 9 on page 1448

NADJ=*value*
specifies the adjustment value that is added to the sample size in the calculation of theoretical percentiles. For additional information, see Chambers et al. (1983)

Default: $\frac{1}{4}$ as recommended by Blom (1958)

NAME='*string*'
specifies a name for the plot, up to eight characters long, that appears in the PROC GREPLAY master menu.

Default: UNIVAR

NCOLS=*n*
specifies the number of columns in the comparative probability plot.

Alias: NCOL=

Default: NCOLS=1, if you specify only one class variable, and NCOLS=2, if you specify two class variables.

Requirement: This option is ignored unless you specify the CLASS statement.

Interaction: If you specify two class variables, you can use the NCOLS= option with the NROWS= option.

NOFRAME
suppresses the frame around the area that is bounded by the axes.

NORMAL<(*normal-suboptions*)>
displays a normal probability plot. This is the default if you omit a distribution option.

Tip: To assess the point pattern, add a diagonal distribution reference line that corresponds to μ_0 and σ_0 with the MU= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of μ_0 and σ_0 with the THETA=EST and SIGMA=EST; the estimates of the mean $\mu_0$and the standard deviation σ_0 are the sample mean and sample standard deviation.

Agreement between the reference line and the point pattern indicates that the normal distribution with parameters μ_0 and σ_0 is a good fit.

Main discussion: "Normal Distribution" on page 1413

See also: the MU= suboption on page 1371 and SIGMA= suboption on page 1373

Featured in: Example 9 on page 1448

NROWS=*n*
specifies the number of rows in the comparative probability plot.

Alias: NROW=

Default: 2

Requirement: This option is ignored unless you specify the CLASS statement.

Interaction: If you specify two class variables, you can use the NCOLS= option with the NROWS= option.

PCTLMINOR
requests minor tick marks for the percentile axis.

Featured in: Example 9 on page 1448

PCTLORDER=*value(s)*
specifies the tick marks that are labeled on the theoretical percentile axis.

Default: 1, 5, 10, 25, 50, 75, 90, 95, and 99

Range: $0 \leq value \leq 100$

Restriction: The values that you specify must be in increasing order and cover the plotted percentile range. Otherwise, PROC UNIVARIATE uses the default.

RANKADJ=*value*
specifies the adjustment value that PROC UNIVARIATE adds to the ranks in the calculation of theoretical percentiles. For additional information, see Chambers et al. (1983).

Default: $-\frac{3}{8}$ as recommended by Blom (1958)

SCALE=*value*
is an alias for the SIGMA= option when you request probability plots with the BETA, EXPONENTIAL, GAMMA, and WEIBULL options and for the ZETA= option when you request the LOGNORMAL option.

See also: the SIGMA= suboption on page 1373 and ZETA= suboption on page 1375

SHAPE=*value(s)*|EST
is an alias for the ALPHA=option when you request gamma plots with the GAMMA option, for the SIGMA= option when you request lognormal plots with the LOGNORMAL option, and for the C= option when you request Weibull plots with the WEIBULL and WEIBULL2 options.

See also: the ALPHA= suboption on page 1367, SIGMA= suboption on page 1373, and C= suboption on page 1368

SIGMA=*value(s)*|EST
specifies the parameter σ, where $\sigma > 0$. The interpretation and use of the SIGMA= option depend on which distribution you specify, as shown Table 41.4 on page 1373.

Table 41.4 Uses of the SIGMA Suboption

Distribution Option	Uses of the SIGMA= Option
BETA, EXPONENTIAL GAMMA, WEIBULL	THETA=θ_0 and SIGMA=σ_0 request a distribution reference line that corresponds to θ_0 and σ_0.
LOGNORMAL	SIGMA=$\sigma_1 \ldots \sigma_n$ requests n probability plots with shape parameters $\sigma_1 \ldots \sigma_n$. The SIGMA= option is required.
NORMAL	MU=μ_0 and SIGMA=σ_0 request a distribution reference line that corresponds to μ_0 and σ_0. SIGMA=EST requests a line with σ_0 equal to the sample standard deviation.
WEIBULL2	SIGMA=σ_0 and C=c_0 request a distribution reference line that corresponds to σ_0 and c_0.

Requirement: You must enclose this suboption in parentheses after the distribution option.

Tip: To compute a maximum likelihood estimate for σ_0, specify SIGMA=EST.

Featured in: Example 9 on page 1448

SLOPE=*value*|EST
specifies the slope for a distribution reference when you request the LOGNORMAL option or WEIBULL2 option.

Requirement: You must enclose this suboption in parentheses after the distribution option.

Tip: When you use the LOGNORMAL option and SLOPE= to request the line, you must also specify a threshold parameter value θ_0 with the THETA= suboption. SLOPE= is an alternative to the ZETA= suboption for specifying ζ_0, because the slope is equal to $\exp(\zeta_0)$.

When you use the WEIBULL2 option and SLOPE= option to request the line, you must also specify a scale parameter value σ_0 with the SIGMA= suboption. SLOPE= is an alternative to the C= suboption for specifying c_0, because the slope is equal to $\frac{1}{c_0}$.

For example, the first and second PROBPLOT statements produce the same probability plots as the third and fourth PROBPLOT statements:

```
proc univariate data=measures;
   probplot width /lognormal(sigma=2 theta=0 zeta=0);
   probplot width /lognormal(sigma=2 theta=0 slope=1);
   probplot width /weibull2(sigma=2 theta=0 c=.25);
   probplot width /weibull2(sigma=2 theta=0 slope=4);
```

Main Discussion: "Three-Parameter Weibull Distribution" on page 1413

SQUARE
displays the probability plot in a square frame.

Default: rectangular frame

THETA=*value*|EST
specifies the lower threshold parameter θ for probability plots when you request the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, WEIBULL, or WEIBULL2 option.

Default: 0

Requirement: You must enclose this suboption in parentheses after the distribution option.

Interaction: When you use the WEIBULL2 option, the THETA= suboption specifies the known lower threshold θ_0, which by default is 0.

When you use the THETA= suboption with another distribution option, THETA= specifies θ_0 for a distribution reference line. To compute a maximum likelihood estimate for θ_0, specify THETA=EST. To request the line, you must also specify a scale parameter.

THRESHOLD= *value*
is an alias for the THETA= option. See the THETA= suboption on page 1374.

VMINOR=*n*
specifies the number of minor tick marks between each major tick mark on the vertical axis. PROBPLOT does not label minor tick marks.

Alias: VM=

Default: 0

VREF=*value(s)*
draws reference lines that are perpendicular to the vertical axis at the *value(s)* that you specify.

See also: CVREF= option on page 1369 and LVREF= option on page 1371.

VREFLABELS=' *label1*'... '*labeln*'
specifies labels for the reference lines that you request with the VREF= option.

Alias: VREFLABEL= and VREFLAB=

Restriction: The number of labels must equal the number of reference lines. Labels can have up to 16 characters.

W=*n*
specifies the width in pixels for a diagonal distribution line.

Default: 1

Requirement: You must enclose this suboption in parentheses after the distribution option.

WEIBULL(C=*value(s)*|EST <*Weibull-suboptions*>)
creates a three-parameter Weibull probability plot for each value of the required shape parameter c.

Alias: WEIB

Requirement: You must specify the shape parameter with the C= suboption.

Interaction: To create a plot that is based on a maximum likelihood estimate for c, specify C=EST.

Tip: To obtain a graphical estimate of c, specify a list of values in the C= suboption. Then select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to θ_0 and σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of θ_0 and σ_0 with THETA=EST and SIGMA=EST.

Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters c, θ_0, and σ_0 is a good fit.

Main discussion: "Three-Parameter Weibull Distribution" on page 1413

See also the C= suboption on page 1368, SIGMA= suboption on page 1373, and THETA= suboption on page 1374

WEIBULL2<(*Weibull-suboptions*)>

creates a two-parameter Weibull probability plot. Use this distribution when your data have a *known* lower threshold θ_0, which by default is 0. To specify the threshold value θ_0, use the THETA= suboption.

Alias: W2

Tip: An advantage of the two-parameter Weibull plot over the three-parameter Weibull plot is that the parameters c and σ can be estimated from the slope and intercept of the point pattern. A disadvantage is that the two-parameter Weibull distribution applies only in situations where the threshold parameter is known.

Tip: To obtain a graphical estimate of c, specify a list of values for the C= suboption. Then select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to σ_0 and c_0 with the SIGMA= and C= suboptions. Alternatively, you can add a distribution reference line that corresponds to estimated values of σ_0 and c_0 with SIGMA=EST and C=EST.

Agreement between the reference line and the point pattern indicates that the Weibull2 distribution with parameters c_0, θ_0, and σ_0 is a good fit.

Main discussion: "Two-Parameter Weibull Distribution" on page 1414

See also: the C= suboption on page 1368, SIGMA= suboption on page 1373, SLOPE= suboption on page 1373, and THETA= suboption on page 1374

ZETA= *value*|EST

specifies a value for the scale parameter ζ for the lognormal probability plots when you request the LOGNORMAL option.

Requirement: You must enclose this suboption in parentheses after the LOGNORMAL option.

Interaction: To request a distribution reference line with intercept θ_0 and slope $\exp(\zeta_0)$, specify THETA= θ_0 and ZETA= ζ_0.

QQPLOT Statement

Creates a quantile-quantile plot (Q-Q plot) (using high-resolution graphics) compares ordered variable values with quantiles of a specified theoretical distribution.

Alias: QQ

Default: Normal Q–Q plot

Restriction: You can specify only one theoretical distribution.

Tip: You can use multiple QQPLOT statements.

Main Discussion: "Quantile–Quantile and Probability Plots" on page 1391

QQPLOT <*variable(s)*> </ *option(s)*>;

To do this:	Use this option:
Request a distribution	
Specify beta probability plot with required shape parameters α, β.	BETA(*beta-suboptions)*
Specify exponential probability plot	EXPONENTIAL(*exponential-suboptions)*
Specify gamma probability plot with a required shape parameter α	GAMMA(*gamma-suboptions*)
Specify lognormal probability plot with a required shape parameter σ	LOGNORMAL(*lognormal-suboptions*)
Specify normal probability plot	NORMAL(*normal-suboptions)*
Specify three-parameter Weibull probability plot with a required shape parameter c	WEIBULL(*Weibull-suboptions*)
Specify two-parameter Weibull probability plot	WEIBULL2(*Weibull2-suboptions*)
Distribution suboptions	
Specify shape parameter α for the beta or gamma distribution	ALPHA=
Specify shape parameter β for the beta distribution	BETA=
Specify shape parameter c for the Weibull distribution or c_0 for distribution reference line of the Weibull2 distribution	C=
Specify μ_0 for distribution reference line of the normal distribution	MU=
Specify σ_0 for distribution reference line for the beta, exponential, gamma, normal, Weibull, or Weibull2 distribution or the required shape parameter σ for the lognormal option	SIGMA=
Specify slope of distribution reference line for the lognormal or Weibull2 distribution	SLOPE=

To do this:	Use this option:
Specify θ_0 for distribution reference line for the beta, exponential, gamma. lognormal, or Weibull distribution, or the lower known threshold θ_0 for the Weibull2 distribution	THETA=
Specify ζ_0 for distribution reference line for the lognormal distribution	ZETA=
Control appearance of distribution reference line	
Specify color of distribution reference line	COLOR=
Specify line type of distribution reference line	L=
Specify width of distribution reference line	W=
Control general plot layout	
Specify reference lines perpendicular to the horizontal axis	HREF=
Specify labels for HREF lines	HREFLABELS=
Adjust sample size when computing quantiles	NADJ=
Suppress frame around plotting area	NOFRAME
Request minor tick marks for percentile axis	PCTLMINOR
Replace theoretical quantiles with percentiles	PCTLSCALE
Adjust ranks when computing quantiles	RANKADJ=
Display Q-Q plot in square format	SQUARE
Specify reference lines perpendicular to the vertical axis	VREF=
Specify labels for VREF lines	VREFLABELS=
Enhance the Q-Q plot	
Specify annotate data set	ANNOTATE=
Specify color for axis	CAXIS=
Specify color for frame	CFRAME=
Specify color for HREF= lines	CHREF=
Specify color for text	CTEXT=
Specify color for VREF= lines	CVREF=
Specify description for plot in graphics catalog	DESCRIPTION=
Specify software font for text	FONT=
Specify number of minor tick marks on horizontal axis	HMINOR=
Specify line style for HREF= lines	LHREF=
Specify line style for VREF= lines	LVREF=
Specify name for plot in graphics catalog	NAME=
Specify number of minor tick marks on vertical axis	VMINOR=
Enhance the comparative Q-Q plot	

To do this:	Use this option:
Apply annotation requested in ANNOTATE= data set to key cell only	ANNOKEY
Specify color for filling frame for row labels	CFRAMESIDE=
Specify color for filling frame for column labels	CFRAMETOP=
Specify distance between tiles	INTERTILE=
Specify number of columns in comparative Q-Q plot	NCOLS=
Specify number of rows in comparative Q-Q plot	NROWS=

Arguments

variable(s)
identifies one or more variables that the procedure uses to create Q-Q plots.

Default: If you omit *variable(s)* in the QQPLOT statement, then the procedure creates a Q-Q plot for each variable that you list in the VAR statement, or for each numeric variable in the DATA= data set if you omit a VAR statement.

Requirement: If you specify a VAR statement, use the *variable(s)* that you list in the VAR statement. Otherwise, *variable(s)* are any numeric variables in the DATA= data set.

Options

ALPHA=*value(s)*|EST
specifies the required shape parameter $\alpha(\alpha > 0)$ for quantile plots when you request the BETA or GAMMA options. The QQPLOT statement creates a plot for each value that you specify.

Requirement: Enclose this suboption in parentheses when it follows the BETA or GAMMA options.

Tip: To compute a maximum likelihood estimate for α, specify ALPHA=EST.

ANNOKEY
specifies to apply the annotation that you requested with the ANNOTATE= option to the *key cell* only. By default, PROC UNIVARIATE applies annotation to all of the cells.

Requirement: This option is ignored unless you specify the CLASS statement.

Tip: Use the KEYLEVEL= option in the CLASS statement to specify the key cell.

See also: the KEYLEVEL= option on page 1334

ANNOTATE=*SAS-data-set*
specifies an input data set that contains annotate variables as described in *SAS/GRAPH Software: Reference*.

Alias: ANNO=

Tip: The ANNOTATE = data set that you specify in the QQPLOT statement is used by all plots that this statement creates. You can also specify an ANNOTATE= data set in the PROC UNIVARIATE statement to enhance all the graphic displays that the procedure creates.

See also: ANNOTATE= on page 1327 in the PROC UNIVARIATE statement

BETA(ALPHA=*value(s)*|EST BETA=*value(s)*|EST <*beta-suboptions*>)
displays a beta Q-Q plot for each combination of the required shape parameters α and β.

Requirement: You must specify the shape parameters with the ALPHA= and BETA= suboptions

Interaction: To create a plot that is based on maximum likelihood estimates for α and β, specify ALPHA=EST and BETA=EST.

Tip: To obtain graphical estimates of α and β, specify lists of values in the ALPHA= and BETA= suboptions. Then select the combination of α and β that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to the lower threshold parameter θ_0 and the scale parameter σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of lower threshold parameter θ_0 and σ_0 with THETA=EST and SIGMA=EST.

Agreement between the reference line and the point pattern indicates that the beta distribution with parameters α, β, θ_0, and σ_0 is a good fit.

Main discussion: "Beta Distribution" on page 1412

See also: the ALPHA= suboption on page 1378, BETA= suboption on page 1379, SIGMA= suboption on page 1383, and THETA= suboption on page 1384.

BETA=*value(s)*|EST
specifies the shape parameter $\beta\,(\beta > 0)$ for Q-Q plots when you request the BETA distribution option. PROC UNIVARIATE creates a plot for each value that you specify.

Alias: B=

Requirement: You must enclose this suboption in parentheses after the BETA option.

Tip: To compute a maximum likelihood estimate for β, specify BETA=EST.

C=*value(s)*|EST
specifies the shape parameter $c\,(c > 0)$ for Q-Q plots when you request the WEIBULL option or WEIBULL2 option. C= is a required suboption in the WEIBULL option.

Requirement: Enclose this suboption in parentheses after the WEIBULL option or WEIBULL2 option.

Interaction: To request a distribution reference line in the WEIBULL2 option, you must specify both the C= and SIGMA= suboptions.

Tip: To compute a maximum likelihood estimate for c, specify C=EST.

CAXIS=*color*
specifies the color for the axes.

Alias: CAXES=

Default: the first color in the device color list

Interaction: This option overrides any COLOR= specification.

CFRAME=*color*
specifies the color for the area that is enclosed by the axes and frame.

Default: the area is not filled

CFRAMESIDE=*color*
specifies the color to fill the frame area for the row labels that display along the left side of the comparative probability plot. This color also fills the frame area for the label of the corresponding class variable (if you associate a label with the variable).

Default: These areas are not filled.

Requirement: This option is ignored unless you specify the CLASS statement.

CFRAMETOP=*color*
specifies the color to fill the frame area for the column labels that display across the top of the comparative probability plot. This color also fills the frame area for the label of the corresponding class variable (if you associate a label with the variable).

Default: These areas are not filled.

Requirement: This option is ignored unless you specify the CLASS statement.

CHREF=*color*
specifies the color for horizontal axis reference lines when you specify the HREF= option.

Default: the first color in the device color list

COLOR=*color*
specifies the color for a distribution reference line.

Default: the fourth color in the device color list

Requirement: You must enclose this suboption in parentheses after a distribution option keyword.

CTEXT=*color*
specifies the color for tick mark values and axis labels.

Default: the color that you specify for the CTEXT= option in the GOPTIONS statement. If you omit the GOPTIONS statement, the default is the first color in the device color list.

CVREF=*color*
specifies the color for the reference lines that you request with the VREF= option.

Alias: CV=

Default: the first color in the device color list.

DESCRIPTION='*string*'
specifies a description, up to 40 characters long, that appears in the PROC GREPLAY master menu.

Alias: DES=

Default: the variable name

EXPONENTIAL<(*exponential-suboptions*)>
displays an exponential Q–Q plot.

Alias: EXP

Tip: To assess the point pattern, add a diagonal distribution reference line that corresponds to θ_0 and σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of the threshold parameter θ_0 and the scale parameter σ_0 with the THETA=EST and SIGMA=EST suboptions.

Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters θ_0 and σ_0 is a good fit.

Main discussion: "Exponential Distribution" on page 1412

See also: the SIGMA suboption on page 1383 and THETA suboption on page 1384

FONT=*font*
specifies a software font for the reference lines and the axis labels.

Default: hardware characters

Interaction: FONT=*font* takes precedence over FTEXT=*font* that you specify in the GOPTIONS statement.

GAMMA(ALPHA=*value(s)*|EST <*gamma-suboptions*>)
displays a gamma Q–Q plot for each value of the required shape parameter α.

Requirement: You must specify the shape parameter with the ALPHA= suboption.

Interaction: To create a plot that is based on a maximum likelihood estimate for α, specify ALPHA=EST.

Tip: To obtain a graphical estimate of α, specify a list of values in the ALPHA= suboption. Then select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to θ_0 and σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of the threshold parameter θ_0 and the scale parameter σ_0 with THETA=EST and SIGMA=EST.

Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters α, θ_0, and σ_0 is a good fit.

Main discussion: "Gamma Distribution" on page 1412

See also: the ALPHA= suboption on page 1378, SIGMA= suboption on page 1383, and THETA= suboption on page 1384

HMINOR=*n*
specifies the number of minor tick marks between each major tick mark on the horizontal axis. PROC UNIVARIATE does not label minor tick marks.

Alias: HM=

Default: 0

HREF=*value(s)*
draws reference lines that are perpendicular to the horizontal axis at the values you specify.

See also: CHREF= option on page 1380

HREFLABELS='*label1*' ... '*labeln*'
specifies labels for the reference lines that you request with the HREF= option.

Alias: HREFLABEL= and HREFLAB=

Restriction: The number of labels must equal the number of reference lines. Labels can have up to 16 characters.

INTERTILE=*value*
specifies the distance in horizontal percentage screen units between the framed areas, which are called *tiles*.

Default: The tiles are contiguous.

Requirement: This option is ignored unless you specify the CLASS statement.

L=*linetype*
specifies the line type for a diagonal distribution reference line.

Default: 1, which produces a solid line

Requirement: You must enclose this suboption in parentheses after a distribution option keyword.

LHREF=*linetype*
specifies the line type for the reference lines that you request with the HREF= option.

Alias: LH=

Default: 2, which produces a dashed line

LOGNORMAL(SIGMA=*value(s)*|EST <*lognormal-suboptions*>)
displays a lognormal Q-Q plot for each value of the required shape parameter σ.

Alias: LNORM

Requirement: You must specify the shape parameter with the SIGMA= suboption.

Tip: To obtain a graphical estimate of σ, specify a list of values for the SIGMA= suboption, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to the threshold parameter θ_0 and the scale parameter ζ_0 with the THETA= and ZETA= suboptions. Alternatively, you can add a line that corresponds to estimated values of θ_0 and ζ_0 with THETA=EST and ZETA=EST. This line has intercept θ_0, and slope $\exp(\zeta_0)$.

Agreement between the reference line and the point pattern indicates that the lognormal distribution with parameters σ, θ_0 and ζ_0 is a good fit.

Main discussion: "Lognormal Distribution" on page 1413

See also: the SIGMA= suboption on page 1383, SLOPE= suboption on page 1384, THETA= suboption on page 1384, and ZETA= suboption on page 1386

LVREF=*linetype*

specifies the line type for the reference lines that you request with the VREF= option.

Alias: LV=

Default: 2, which produces a dashed line

MU=*value*|EST

specifies the mean μ for a normal Q-Q plot requested with the NORMAL option.

Default: the sample mean

Requirement: You must enclose this suboption in parentheses after the NORMAL option.

Tip: Specify the MU= and SIGMA= suboptions together to request a distribution reference line. Specify MU=EST to request a distribution reference line with μ_0 equal to the sample mean.

NADJ=*value*

specifies the adjustment value that is added to the sample size in the calculation of theoretical quantiles. For additional information, see Chambers et al. (1983).

Default: $\frac{1}{4}$ as recommended by Blom (1958)

NAME='*string*'

specifies a name for the plot, up to eight characters long, that appears in the PROC GREPLAY master menu.

Default: UNIVAR

NCOLS=*n*

specifies the number of columns in the comparative probability plot.

Alias: NCOL=

Default: NCOLS=1, if you specify only one class variable, and NCOLS=2, if you specify two class variables.

Requirement: This option is ignored unless you specify the CLASS statement.

Interaction: If you specify two class variables, you can use the NCOLS= option with the NROWS= option.

NOFRAME

suppresses the frame around the area that is bounded by the axes.

NORMAL<(*normal-suboptions*)>

displays a normal Q-Q plot. This is the default if you omit a distribution option.

Tip: To assess the point pattern, add a diagonal distribution reference line that corresponds to μ_0 and σ_0 with the MU= and SIGMA= suboptions. Alternatively,

you can add a line that corresponds to estimated values of μ_0 and σ_0 with the THETA=EST and SIGMA=EST; the estimates of the mean μ_0 and the standard deviation σ_0 are the sample mean and sample standard deviation.

Agreement between the reference line and the point pattern indicates that the normal distribution with parameters μ_0 and σ_0 is a good fit.

Main discussion: "Normal Distribution" on page 1413

See also: the MU= suboption on page 1382 and SIGMA= suboption on page 1383

NROWS=*n*

specifies the number of rows in the comparative probability plot.

Alias: NROW=

Default: 2

Requirement: This option is ignored unless you specify the CLASS statement.

Interaction: If you specify two class variables, you can use the NCOLS= option with the NROWS= option.

PCTLMINOR

requests minor tick marks for the percentile axis.

PCTLSCALE

requests scale labels for the theoretical quantile axis in percentile units, resulting in a nonlinear axis scale.

Tip: Tick marks are drawn uniformly across the axis based on the quantile scale. In all other respects, the plot remains the same, and you must specify HREF= values in quantile units. For a true nonlinear axis, use the PROBPLOT statement.

RANKADJ=*value*

specifies the adjustment value that PROC UNIVARIATE adds to the ranks in the calculation of theoretical quantiles. For additional information, see Chambers et al. (1983).

Default: $-\frac{3}{8}$ as recommended by Blom (1958)

SCALE=*value*

is an alias for the SIGMA= option when you request Q-Q plots with the BETA, EXPONENTIAL, GAMMA, WEIBULL, and WEIBULL2 options and for the ZETA= option when you request the LOGNORMAL option.

See also: SIGMA= on page 1383 and ZETA= on page 1386

SHAPE=*value(s)*|EST

is an alias for the ALPHA=option when you request gamma plots with the GAMMA option, for the SIGMA= option when you request lognormal plots with the LOGNORMAL option, and for the C= option when you request Weibull plots with the WEIBULL, and WEIBULL2 options.

See also: ALPHA= on page 1378, SIGMA= on page 1383, and C= on page 1379

SIGMA=*value(s)*|EST

specifies the distribution parameter σ, where $\sigma > 0$ for the quantile plot. The interpretation and use of the SIGMA= option depend on which distribution you specify, as shown in Table 41.5 on page 1384.

Table 41.5 Uses of the SIGMA Suboption

Distribution Option	Uses of the SIGMA= Option
BETA, EXPONENTIAL GAMMA, WEIBULL	THETA=θ_0 and SIGMA=σ_0 request a distribution reference line with intercept θ_0 and slope σ_0.
LOGNORMAL	SIGMA=$\sigma_1 \ldots \sigma_n$ requests n Q-Q plots with shape parameters $\sigma_1 \ldots \sigma_n$. The SIGMA= option is required.
NORMAL	MU=μ_0 and SIGMA=σ_0 request a distribution reference line with intercept μ_0 and slope σ_0. SIGMA=EST requests a slope σ_0 equal to the sample standard deviation.
WEIBULL2	SIGMA=σ_0 and C=c_0 request a distribution reference line with intercept $\log(\sigma_0)$ and slope $\frac{1}{c_0}$.

Requirement: Enclose this suboption in parentheses after the distribution option.

Tip: To compute a maximum likelihood estimate for σ_0, specify SIGMA=EST .

SLOPE=*value*|EST

specifies the slope for a distribution reference when you request the LOGNORMAL option or WEIBULL2 option.

Requirement: Enclose this suboption in parentheses after the distribution option.

Tip: When you use the LOGNORMAL option and SLOPE= to request the line, you must also specify a threshold parameter value θ_0 with the THETA= suboption. SLOPE= is an alternative to the ZETA= suboption for specifying ζ_0, because the slope is equal to $\exp(\zeta_0)$.

When you use the WEIBULL2 option and SLOPE= option to request the line, you must also specify a scale parameter value σ_0 with the SIGMA= suboption. SLOPE= is an alternative to the C= suboption for specifying c_0, because the slope is equal to $\frac{1}{c_0}$.

For example, the first and second QQPLOT statements produce the same quantile-quantile plots as the third and fourth QQPLOT statements:

```
proc univariate data=measures;
   qqplot width /lognormal(sigma=2 theta=0 zeta=0);
   qqplot width /lognormal(sigma=2 theta=0 slope=1);
   qqplot width /weibull2(sigma=2 theta=0 c=.25);
   qqplot width /weibull2(sigma=2 theta=0 slope=4);
```

Main Discussion: "Shape Parameters" on page 1414

SQUARE

displays the Q-Q plot in a square frame.

Default: rectangular frame

THETA=*value*|EST

specifies the lower threshold parameter θ for Q-Q plots when you request BETA, EXPONENTIAL, GAMMA, LOGNORMAL, WEIBULL, or WEIBULL2 option.

Default: 0

Requirement: You must enclose this suboption in parentheses after the distribution option.

Interaction: When you use the WEIBULL2 option, the THETA= suboption specifies the known lower threshold θ_0, which by default is 0.

When you use the THETA= suboption with another distribution option, THETA= specifies θ_0 for a distribution reference line. To compute a maximum likelihood estimate for θ_0, specify THETA=EST. To request the line, you must also specify a scale parameter.

THRESHOLD= *value*|EST
is an alias for the THETA= option. See the THETA= suboption on page 1384.

VMINOR=*n*
specifies the number of minor tick marks between each major tick mark on the vertical axis. QQPLOT does not label minor tick marks.

Alias: VM=

Default: 0

VREF=*value(s)*
draws reference lines that are perpendicular to the vertical axis at the *value(s)* you specify.

See also: CVREF= option on page 1380 and LVREF= option on page 1382

VREFLABELS=' *label1*'... '*labeln*'
specifies labels for the reference lines that you request with the VREF= option.

Alias: VREFLABEL= and VREFLAB=

Restriction: The number of labels must equal the number of reference lines. Labels can have up to 16 characters.

W=*n*
specifies the width in pixels for a distribution reference line.

Default: 1

Requirement: You must enclose this suboption in parentheses after the distribution option.

WEIBULL(C=*value(s)*|EST <*Weibull-suboptions*>)
creates a three-parameter Weibull Q–Q plot for each value of the required shape parameter c.

Alias: WEIB

Requirement: You must specify the shape parameter with the C= suboption.

Interaction: To create a plot that is based on a maximum likelihood estimate for c, specify C=EST.

To specify the threshold value θ_0, use the THETA= suboption.

Tip: To obtain a graphical estimate of c, specify a list of values in the C= suboption. Then select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line with intercept θ_0 and slope σ_0 with the THETA= and SIGMA= suboptions. Alternatively, you can add a line that corresponds to estimated values of θ_0 and σ_0 with THETA=EST and SIGMA=EST.

Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters c, θ_0, and σ_0 is a good fit.

Main discussion: "Three-Parameter Weibull Distribution" on page 1413

See also the C= suboption on page 1379, SIGMA= suboption on page 1383, and THETA= suboption on page 1384

WEIBULL2<(*Weibull-suboptions*)>
creates a two-parameter Weibull Q-Q plot. Use this distribution when your data have a *known* lower threshold θ_0, which by default is 0. To specify the threshold value θ_0, use the THETA= suboption.

Note: The C= shape parameter option is not required with the Weibull2 option. △

Alias: W2

Default: 0

Interaction: To specify the threshold value θ_0, use the THETA= suboption.

Tip: An advantage of the two-parameter Weibull plot over the three-parameter Weibull plot is that the parameters c and σ can be estimated from the slope and intercept of the point pattern. A disadvantage is that the two-parameter Weibull distribution applies only in situations where the threshold parameter is known.

Tip: To obtain a graphical estimate of θ_0, specify a list of values for the THETA= suboption. Then select the value that most nearly linearizes the point pattern.

To assess the point pattern, add a diagonal distribution reference line that corresponds to σ_0 and c_0 with the SIGMA= and C= suboptions. Alternatively, you can add a distribution reference line that corresponds to estimated values of σ_0 and c_0 with SIGMA=EST and C=EST.

Agreement between the reference line and the point pattern indicates that the Weibull2 distribution with parameters c_0, θ_0, and σ_0 is a good fit.

Main discussion: "Two-Parameter Weibull Distribution" on page 1414

See also: the C= suboption on page 1379, SIGMA= suboption on page 1383, SLOPE= suboption on page 1384, and THETA= suboption on page 1384

ZETA= *value*|EST
specifies a value for the scale parameter ζ for the lognormal Q-Q plots when you request the LOGNORMAL option.

Requirement: You must enclose this suboption in parentheses after the LOGNORMAL option.

Interaction: To request a distribution reference line with intercept θ_0 and slope $\exp(\zeta_0)$, specify THETA= θ_0 and ZETA= ζ_0.

Theoretical Percentiles of Quantile-Quantile Plots

To estimate percentiles from a Q-Q plot

- Specify the PCTLAXIS option, which adds a percentile axis opposite the theoretical quantile axis. The scale for the percentile axis ranges between 0 and 100 with tick marks at percentile values such as 1, 5, 10, 25, 50, 75, 90, 95, and 99.
- Specify the PCTLSCALE option, which relabels the horizontal axis tick marks with their percentile equivalents but does not alter their spacing. For example, on a normal Q-Q plot, the tick mark labeled 0 is relabeled as 50 because the 50^{th} percentile corresponds to the zero quantile.

You can also use the PROBPLOT statement to estimate percentiles.

VAR Statement

Specifies the analysis variables and their order in the results.

Default: If you omit the VAR statement, PROC UNIVARIATE analyzes all numeric variables that are not listed in the other statements.

Featured in: Example 1 on page 1418 and Example 6 on page 1437

VAR *variable(s)*;

Required Arguments

variable(s)
identifies one or more analysis variables.

Using the Output Statement with the VAR Statement

Use a VAR statement when you use an OUTPUT statement. To store the same statistic for several analysis variables in the OUT= data set, you specify a list of names in the OUTPUT statement. PROC UNIVARIATE makes a one-to-one correspondence between the order of the analysis variables in the VAR statement and the list of names that follow a statistic keyword.

WEIGHT Statement

Specifies weights for analysis variables in the statistical calculations.

See also: For information about how to calculate weighted statistics and for an example that uses the WEIGHT statement, see "Calculating Weighted Statistics" on page 74

WEIGHT *variable*;

Required Arguments

variable
specifies a numeric variable whose values weight the values of the analysis variables. The values of the variable do not have to be integers. If the value of the weight variable is

Weight value…	PROC UNIVARIATE…
0	counts the observation in the total number of observations
less than 0	converts the weight value to zero and counts the observation in the total number of observations
missing	excludes the observation

To exclude observations that contain negative and zero weights from the analysis, use EXCLNPWGT. Note that most SAS/STAT procedures, such as PROC GLM, exclude negative and zero weights by default.

The weight variable does not change how the procedure determines the range, mode, extreme values, extreme observations, or number of missing values. The Student's t test is the only test of location that PROC UNIVARIATE computes when you weight the analysis variables.

Restriction: The CIPCTLDF, CIPCTLNORMAL, LOCCOUNT, NORMAL, ROBUSTSCALE, TRIMMED=, and WINSORIZED= options are not available with the WEIGHT statement.

Restriction: To compute weighted skewness or kurtosis, use VARDEF=DF or VARDEF=N in the PROC statement.

Tip: When you use the WEIGHT statement, consider which value of the VARDEF= option is appropriate. See VARDEF= on page 1331 and the calculation of weighted statistics in "Keywords and Formulas" on page 1458 for more information.

Note: Prior to Version 7 of the SAS System, the procedure did not exclude the observations with missing weights from the count of observations. △

Concepts

Rounding

When you specify ROUND=*u*, PROC UNIVARIATE rounds a variable by using the rounding unit to divide the number line into intervals with midpoints *u***i*, where *u* is the nonnegative rounding unit and *i* equals the integers (..., -4, -3, -2, -1, 0, 1, 2, 3, 4,...). The interval width is *u*. Any variable value that falls in an interval rounds to the midpoint of that interval. A variable value that is midway between two midpoints, and is therefore on the boundary of two intervals, rounds to the even midpoint. Even midpoints occur when *i* is an even integer (0,±2,±4,...).

When ROUND=1 and the analysis variable values are between -2.5 and 2.5, the intervals are as follows:

i	Interval	Midpoint	Left endpt rounds to	Right endpt rounds to
-2	[-2.5,-1.5]	-2	-2	-2
-1	[-1.5,-0.5]	-1	-2	0
0	[-0.5,0.5]	0	0	0
1	[0.5,1.5]	1	0	2
2	[1.5,2.5]	2	2	2

When ROUND=.5 and the analysis variable values are between -1.25 and 1.25, the intervals are as follows:

i	Interval	Midpoint	Left endpt rounds to	Right endpt rounds to
-2	[-1.25,-0.75]	-1.0	-1	-1
-1	[-0.75,-0.25]	-0.5	-1	0
0	[-0.25,0.25]	0.0	0	0
1	[0.25,0.75]	0.5	0	1
2	[0.75,1.25]	1.0	1	1

As the rounding unit increases, the interval width also increases. This reduces the number of unique values and decreases the amount of memory that PROC UNIVARIATE needs.

Generating Line Printer Plots

The PLOTS option in the PROC UNIVARIATE statement provides up to four diagnostic line printer plots to examine the data distribution. These plots are the stem-and-leaf plot or horizontal bar chart, the box plot, the normal probability plot, and the side-by-side box plots. If you specify the WEIGHT statement, PROC UNIVARIATE provides a weighted histogram, a weighted box plot based on the weighted quantiles, and a weighted normal probability plot.

Stem-and-Leaf Plot

The first plot in the output is either a stem-and-leaf plot (Tukey 1977) or a horizontal bar chart. If any single interval contains more than 49 observations, the horizontal bar chart appears. Otherwise, the stem-and-leaf plot appears. The stem-and-leaf plot is like a horizontal bar chart in that both plots provide a method to visualize the overall distribution of the data. The stem-and-leaf plot provides more detail because each point in the plot represents an individual data value.

To change the number of stems that the plot displays, use PLOTSIZE= to increase or decrease the number of rows. Instructions that appear below the plot explain how to determine the values of the variable. If no instructions appear, you multiply *Stem.Leaf* by 1 to determine the values of the variable. For example, if the stem value is 10 and the leaf value is 1, then the variable value is approximately 10.1.

For the stem-and-leaf plot, the procedure rounds a variable value to the nearest leaf. If the variable value is exactly halfway between two leaves, the value rounds to the nearest leaf with an even integer value. For example, a variable value of 3.15 has a stem value of 3 and a leaf value of 2.

Box Plot

The box plot, also known as a schematic plot, appears beside the stem-and-leaf plot. Both plots use the same vertical scale. The box plot provides a visual summary of the data and identifies outliers. The bottom and top edges of the box correspond to the sample 25th (Q1) and 75th (Q3) percentiles. The box length is one *interquartile range* (Q3 - Q1). The center horizontal line with asterisk endpoints corresponds to the sample median. The central plus sign (+) corresponds to the sample mean. If the mean and median are equal, the plus sign falls on the line inside the box. The vertical lines that project out from the box, called *whiskers*, extend as far as the data extend, up to a distance of 1.5 interquartile ranges. Values farther away are potential outliers. The procedure identifies the extreme values with a zero or an asterisk (*). If zero appears, the value is between 1.5 and 3 interquartile ranges from the top or bottom edge of the box. If an asterisk appears, the value is more extreme.

To generate box plot using high-resolution graphics, use the BOXPLOT procedure in SAS/STAT software.

Normal Probability Plot

The normal probability plot is a quantile-quantile plot of the data. The procedure plots the empirical quantiles against the quantiles of a standard normal distribution. Asterisks (*) indicate the data values. The plus signs (+) provide a straight reference line that is drawn by using the sample mean and standard deviation. If the data are from a normal distribution, the asterisks tend to fall along the reference line. The vertical coordinate is the data value, and the horizontal coordinate is $\Phi^{-1}(v_i)$ where

$$\Phi^{-1}\left((r_i - 3/8) \,/\, (n + 1/4)\right)$$

and where

v_i	is $\left(r_i - \frac{3}{8}\right) / \left(n + \frac{1}{4}\right)$.
Φ^{-1}	is the inverse of the standard normal distribution function.
r_i	is the rank of the ith data value when ordered from smallest to largest.
n	is the number of nonmissing data values.

For weighted normal probability plot, the *i*th ordered observation is plotted against the normal quantile $\Phi^{-1}(v_i)$, where Φ^{-1} is the inverse standard cumulative normal distribution and

$$v_i = \frac{\sum_{j=1}^{i} w_{(j)} \left(1 - \frac{3}{8i}\right)}{W\left(1 + \frac{1}{4n}\right)}$$

where $w_{(j)}$ is weight that is associated with $y_{(j)}$ for the jth ordered observation and $W = \sum_{i=1}^{n} w_i$ is the sum of the individual weights.

When each observation has an identical weight, $w_{(j)} = w$, the formula for v_i reduces to the expression for v_i in the unweighted normal probability plot

$$v_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

When the value of VARDEF= is WDF or WEIGHT, PROC UNIVARIATE draws a reference line with intercept $\widehat{\mu}$ and slope $\widehat{\sigma}$ and when the value of VARDEF= is DF or N, the slope is $\widehat{\sigma}/\sqrt{\overline{w}}$ where $\overline{w} = W/n$ is the average weight.

When each observation has an identical weight and the value of VARDEF= is DF, N, or WEIGHT, the reference line reduces to the usual reference line with intercept $\widehat{\mu}$ and slope $\widehat{\sigma}$ in the unweighted normal probability plot.

If the data are normally distributed with mean μ, standard deviation σ, and each observation has an identical weight w, then, as in the unweighted normal probability plot, the points on the plot should lie approximately on a straight line. The intercept is μ and slope is σ when VARDEF= is WDF or WEIGHT, and the slope is $\sigma/\sqrt{w}$ when VARDEF= is DF or N.

Side-by-Side Box Plots

When you use a BY statement with the PLOT option, PROC UNIVARIATE produces full-page side-by-side box plots, one for each BY group. The box plots (also known as schematic plots) use a common scale that allows you to compare the data distribution across BY groups. This plot appears after the univariate analyses of all BY groups. Use the NOBYPLOT option to suppress this plot.

For more information on how to interpret these plots see *SAS System for Elementary Statistical Analysis* and *SAS System for Statistical Graphics*.

Generating High-Resolution Graphics

If your site licenses SAS/GRAPH software, you can use the HISTOGRAM statement, PROBPLOT statement, and QQPLOT statement to create high-resolution graphs.

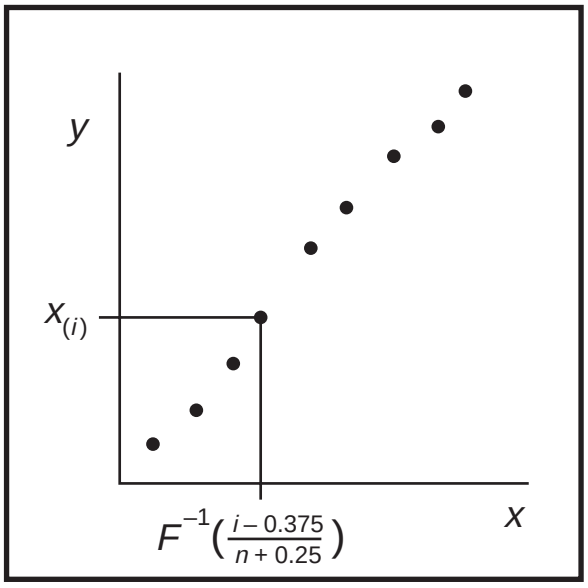
The HISTOGRAM statement generates histograms and comparative histograms that allow you to examine the data distribution. You can optionally fit families of density curves and superimpose kernel density estimates on the histograms. For additional information about the fitted distributions and kernel density estimates, see "Formulas for Fitted Continuous Distributions" on page 1406.

The PROBPLOT statement generates a probability plot, which compares ordered values of a variable with percentiles of a specified theoretical distribution. The QQPLOT statement generates a quantile-quantile plot, which compares ordered values of a variable with quantiles of a specified theoretical distribution. Thus, you can use these plots to determine how well a theoretical distribution models a set of measures.

Quantile–Quantile and Probability Plots

The following figure illustrates how to construct a Q-Q plot for a specified theoretical distribution $F(x)$ with the QQPLOT statement.

Figure 41.3 Construction of a Q-Q Plot



First, the n nonmissing values of the variable are ordered from smallest to largest: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Then, the i^{th} ordered value $x_{(i)}$ is represented on the plot by a point whose y-coordinate is $x_{(i)}$ and whose x-coordinate is $F^{-1}\left(\frac{i-0.375}{n+0.25}\right)$, where $F(.)$ is the theoretical distribution with a zero location parameter and a unit scale parameter. For additional information about the theoretical distributions that you can request, see "Theoretical Distributions for Quantile-Quantile and Probability Plots" on page 1411.

You can modify the adjustment constants -0.375 and 0.25 with the RANKADJ= and NADJ= options. The default combination is recommended by Blom (1958). For additional information, see Chambers et al. (1983). Since $x_{(i)}$ is a quantile of the empirical cumulative distribution function (ecdf), a Q-Q plot compares quantiles of the

ecdf with quantiles of a theoretical distribution. Probability plots are constructed the same way, except that the x-axis is scaled nonlinearly in percentiles.

Interpreting Quantile–Quantile and Probability Plots

If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern. Thus, you can use a Q–Q plot or a probability plot to determine how well a theoretical distribution models a set of measurements. The following properties of these plots make them useful diagnostics to test how well a specified theoretical distribution fits a set of measurements:

- If the quantiles of the theoretical and data distributions agree, the plotted points fall on or near the line $y = x$.
- If the theoretical and data distributions differ only in their location or scale, the points on the plot fall on or near the line $y = ax + b$. The slope a and intercept b are visual estimates of the scale and location parameters of the theoretical distribution.

Q-Q plots are more convenient than probability plots for graphical estimation of the location and scale parameters because the x-axis of a Q-Q plot is scaled linearly. On the other hand, probability plots are more convenient for estimating percentiles or probabilities. There are many reasons why the point pattern in a Q-Q plot may not be linear. Chambers et al. (1983) and Fowlkes (1987) discuss the interpretations of commonly encountered departures from linearity, and these are summarized in the following table.

Table 41.6 Quantile-Quantile Plot Diagnostics

Description of Point Pattern	Possible Interpretation
All but a few points fall on a line	Outliers in the data
Left end of pattern is below the line; right end of pattern is above the line	Long tails at both ends of the data distribution
Left end of pattern is above the line; right end of pattern is below the line	Short tails at both ends of the distribution
Curved pattern with slope increasing from left to right	Data distribution is skewed to the right
Curved pattern with slope decreasing from left to right	Data distribution is skewed to the left
Staircase pattern (plateaus and gaps)	Data have been rounded or are discrete

In some applications, a nonlinear pattern may be more revealing than a linear pattern. However as noted by Chambers et al. (1983), departures from linearity can also be due to chance variation.

Determining Computer Resources

Because PROC UNIVARIATE computes quantile statistics, it requires additional memory to store a copy of the data in memory. By default, the report procedures PROC MEANS, PROC SUMMARY, and PROC TABULATE require less memory because they do not automatically compute quantiles. These procedures also provide an option to use

a new fixed-memory quantiles estimation method that is usually less memory intense. For more information, see "Quantiles" on page 653.

The only factor that limits the number of variables that you can analyze is the computer resources that are available. The amount of temporary storage and CPU time that PROC UNIVARIATE requires depends on the statements and the options that you specify. To calculate the computer resources the procedure needs, let

N be the number of observations in the data set

V be the number of variables in the VAR statement

U_i be the number of unique values for the ith variable.

Then the minimum memory requirement in bytes to process all variables is

$$M = 24 \sum U_i$$

If M bytes are not available, PROC UNIVARIATE must process the data multiple times to compute all the statistics. This reduces the minimum memory requirement to

$$M = 24 \max\,(U_i)$$

ROUND= reduces the number of unique values (U_i), thereby reducing memory requirements. ROBUSTSCALE requires $40U_i$ bytes of temporary storage.

Several factors affect the CPU time requirement:

- The time to create V tree structures to internally store the observations is proportional to $NV \log(N)$.
- The time to compute moments and quantiles for the ith variable is proportional to U_i.
- The time to compute the NORMAL option test statistics is proportional to N.
- The time to compute the ROBUSTSCALE option test statistics is proportional to $U_i \log(U_i)$.
- The time to compute the exact significance level of the sign rank statistic may increase when the number of nonzero values is less than or equal to 20.

Each of these factors has a different constant of proportionality. For additional information on how to optimize CPU performance and memory usage, see the SAS documentation for your operating environment.

Statistical Computations

PROC UNIVARIATE uses standard algorithms to compute the moment statistics (such as the mean, variance, skewness, and kurtosis). See Appendix 1, "SAS Elementary Statistics Procedures," on page 1457 for the statistical formulas. The computational details for confidence limits, hypothesis test statistics, and quantile statistics follow.

Confidence Limits for Parameters of the Normal Distribution

The two-sided $100(1-\alpha)$ percent confidence interval for the mean has upper and lower limits

$$\overline{x} \pm t_{(1-\alpha/2;n-1)} \frac{s}{\sqrt{n}}$$

where s is $\sqrt{\frac{1}{n-1} \sum (x_i - \overline{x})^2}$ and $t_{(1-\alpha/2;n-1)}$ is the $(1 - \alpha/2)$ percentile of the t distribution with $n - 1$ degrees of freedom.

The one-sided $100\,(1 - \alpha)$ percent confidence limit is computed as

$$\overline{x} + t_{(1-\alpha;n-1)} \frac{s}{\sqrt{n}} \qquad \text{(upper)}$$

$$\overline{x} - t_{(1-\alpha;n-1)} \frac{s}{\sqrt{n}} \qquad \text{(lower)}$$

The two-sided $100\,(1 - \alpha)$ percent confidence interval for the standard deviation has lower and upper limits

$$s\sqrt{\frac{n-1}{\chi^2_{(1-\alpha/2;n-1)}}},\ s\sqrt{\frac{n-1}{\chi^2_{(\alpha/2;n-1)}}}$$

where $\chi^2_{(1-\alpha/2;n-1)}$ and $\chi^2_{(\alpha/2;n-1)}$ are the $(1 - \alpha/2)$ and $\alpha/2$ percentiles of the chi-square distribution with $n - 1$ degrees of freedom. A one-sided $100\,(1 - \alpha)$ percent confidence limit is computed by replacing $\alpha/2$ with α.

A $100\,(1 - \alpha)$ percent confidence interval for the variance has upper and lower limits equal to the squares of the corresponding upper and lower limits for the standard deviation.

When you use the WEIGHT statement and specify VARDEF=DF in the PROC statement, the $100\,(1 - \alpha)$ percent confidence interval for the weighted mean is

$$\overline{x}_w \pm t_{(1-\alpha/2)} \frac{s_w}{\sqrt{\sum_{i=1}^{n} w_i}}$$

where $\overline{x}_w$ is the weighted mean, s_w is the weighted standard deviation, w_i is the weight for ith observation, and $t_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ critical percentage for the t distribution with $n - 1$ degrees of freedom.

Tests for Location

PROC UNIVARIATE computes tests for location that include Student's t test, the sign test, and the Wilcoxon signed rank test. All three tests produce a test statistic for the null hypothesis that the mean or median is equal to a given value μ_0 against the two-sided alternative that the mean or median is not equal to μ_0. By default, PROC UNIVARIATE sets the value of μ_0 to zero. Use the MU0= option in the PROC UNIVARIATE statement to test that the mean or median is equal to another value.

The Student's t test is appropriate when the data are from an approximately normal population; otherwise, use nonparametric tests such as the sign test or the signed rank test. For large sample situations, the t test is asymptotically equivalent to a z test.

If you use the WEIGHT statement, PROC UNIVARIATE computes only one weighted test for location, the *t* test. You must use the default value for the VARDEF= option in the PROC statement.

You can also compare means or medians of *paired data*. Data are said to be paired when subjects or units are matched in pairs according to one or more variables, such as pairs of subjects with the same age and gender. Paired data also occur when each subject or unit is measured at two times or under two conditions. To compare the means or medians of the two times, create an analysis variable that is the difference between the two measures. The test that the mean or the median difference of the variables equals zero is equivalent to the test that the means or medians of the two original variables are equal. See Example 4 on page 1428.

Student's *t* Test

PROC UNIVARIATE calculates the *t* statistic as

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\overline{x}$ is the sample mean, n is the number of nonmissing values for a variable, and s is the sample standard deviation. Under the null hypothesis, the population mean equals μ_0. When the data values are approximately normally distributed, the probability under the null hypothesis of a *t* statistic that is as extreme, or more extreme, than the observed value (the *p*-value) is obtained from the *t* distribution with $n - 1$ degrees of freedom. For large n, the *t* statistic is asymptotically equivalent to a *z* test.

When you use the WEIGHT statement and the default value of VARDEF=, which is DF, the *t* statistic is calculated as

$$t_w = \frac{\overline{x}_w - \mu_0}{s_w/\sqrt{\sum_{i=1}^{n} w_i}}$$

where $\overline{x}_w$ is the weighted mean, s_w is the weighted standard deviation, and w_i is the weight for ith observation. The t_w statistic is treated as having a Student's *t* distribution with $n - 1$ degrees of freedom. If you specify the EXCLNPWGT option in the PROC statement, n is the number of nonmissing observations when the value of the WEIGHT variable is positive. By default, n is the number of nonmissing observations for the WEIGHT variable.

Sign Test

PROC UNIVARIATE calculates the sign test statistic as

$$\mathbf{M} = \left(n^+ - n^-\right)/2$$

where n^+ is the number of values that is greater than u_0 and n^- is the number of values that is less than u_0. Values equal to u_0 are discarded.

Under the null hypothesis that the population median is equal to u_0, the *p*-value for the observed statistic *M* is

$$\text{Prob}\left\{|\mathbf{M}| \geq |M|\right\} = 0.5^{(n_t-1)} \sum_{j=0}^{\min(n^+,n^-)} \binom{n_t}{j}$$

where $n_t = n^+ + n^-$ is the number of x_i values not equal to u_0.

Wilcoxon Signed Rank Test

PROC UNIVARIATE calculates the Wilcoxon signed rank test statistic as

$$S = \sum r_i^+ - n_t\left(n_t + 1\right)/4$$

where r_i^+ is the rank of $|x_i - \mu_0|$ after discarding values of x_i equal to u_0, n_t is the number of x_i values not equal to u_0, and the sum is calculated for values of $x_i - \mu_0$ greater than 0. Average ranks are used for tied values.

The *p*-value is the probability of obtaining a signed rank statistic greater in absolute value than the absolute value of the observed statistic *S*. If $n_t \leq 20$, the significance level of S is computed from the exact distribution of S, which can be enumerated under the null hypothesis that the distribution is symmetric about u_0. When $n_t > 20$, the significance of level S is computed by treating

$$\frac{S\sqrt{n_t - 1}}{\sqrt{n_t V - S^2}}$$

as a Student's *t* variate with $n_t - 1$ degrees of freedom. V is computed as

$$V = \frac{1}{24} n_t\left(n_t + 1\right)\left(2n_t + 1\right) - 0.5\sum t_i\left(t_i + 1\right)\left(t_i - 1\right)$$

where the sum is calculated over groups that are tied in absolute value, and t_i is the number of tied values in the ith group (Iman 1974; Conover 1980).

The Wilcoxon signed rank test assumes that the distribution is symmetric. If the assumption is not valid, you can use the sign test to test that the median is u_0. See Lehmann (1975) for more details.

Goodness-of-Fit Tests

When you specify the NORMAL option in the PROC UNIVARIATE statement or you request a fitted parametric distribution in the HISTOGRAM statement, the procedure computes test statistics for the null hypothesis that the values of the analysis variable are a random sample from the specified theoretical distribution. When you specify the normal distribution, the test statistics depend on the sample size. If the sample size is less than or equal to 2000, PROC UNIVARIATE calculates the Shapiro-Wilk *W* statistic. For a specified distribution, the procedure attempts to calculate three goodness-of-fit tests that are based on the empirical distribution function (EDF): the Kolmogorov-Smirnov *D* statistic, the Anderson-Darling statistic, and the Cramer-von Mises statistic. However, some of the EDF tests are currently not supported when the

parameters of a specified distribution are estimated. See Table 41.7 on page 1400 for more information.

You determine whether to reject the null hypothesis by examining the probability that is associated with a test statistic. When the *p*-value is less than the predetermined critical value (alpha value), you reject the null hypothesis and conclude that the data came from the theoretical distribution.

If you want to test the normality assumptions that underlie analysis of variance methods, beware of using a statistical test for normality alone. A test's ability to reject the null hypothesis (known as the *power* of the test) increases with the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. Since small deviations from normality do not severely affect the validity of analysis of variance tests, it is important to examine other statistics and plots to make a final assessment of normality. The skewness and kurtosis measures and the plots that are provided by the PLOTS option, the HISTOGRAM statement, PROBPLOT statement, and QQPLOT statement can be very helpful. For small sample sizes, power is low for detecting larger departures from normality that may be important. To increase the test's ability to detect such deviations, you may want to declare significance at higher levels, such as 0.15 or 0.20, rather than the often-used 0.05 level. Again, consulting plots and additional statistics will help you assess the severity of the deviations from normality.

Shapiro-Wilk Statistic

If the sample size is less than or equal to 2000 and you specify the NORMAL option, PROC UNIVARIATE computes the Shapiro-Wilk statistic, *W*. The *W* statistic is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance (Shapiro, 1965). *W* must be greater than zero and less than or equal to one. Small values of *W* lead to the rejection of the null hypothesis of normality. The distribution of *W* is highly skewed. Seemingly large values of *W* (such as 0.90) may be considered small and lead you to reject the null hypothesis. When the sample size is greater than three, the coefficients to compute the linear combination of the order statistics are approximated by the method of Royston (1992).

$$Z_n = \left(-\log\left(\gamma - \log\left(1 - W_n\right)\right) - \mu\right)/\sigma$$

when $4 \leq n \leq 11$ and

$$Z_n = \left(\log\left(1 - W_n\right) - \mu\right)/\sigma$$

when $12 \leq n \leq 2000$, where γ, μ, and σ are functions of n, obtained from simulation results, and Z_n is a standard normal variate. Large values of Z_n indicate departure from normality.

EDF Goodness-of-Fit Tests

When you fit a parametric distribution, PROC UNIVARIATE provides a series of goodness-of-fit tests that are based on the empirical distribution function (EDF). The empirical distribution function is defined for a set of n independent observations $X_1, \ldots, X_n$ with a common distribution function $F\left(x\right)$. The observations that are

ordered from smallest to largest as $X_{(1)}, \ldots, X_{(n)}$. The empirical distribution function, $F_n(x)$, is defined as

$$
\begin{aligned}
F_n(x) &= 0, \quad x < X_{(1)} \\
F_n(x) &= \frac{i}{n}, \quad X_{(i)} \leq x < X_{(i+1)} \quad i = 1, \ldots, n-1 \\
F_n(x) &= 1, \quad X_{(n)} \leq x
\end{aligned}
$$

Note that $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. This function estimates the distribution function $F(x)$. At any value x, $F_n(x)$ is the proportion of observations that is less than or equal to x while $F(x)$ is the theoretical probability of an observation that is less than or equal to x. EDF statistics measure the discrepancy between $F_n(x)$ and $F(x)$.

The computational formulas for the EDF statistics use the probability integral transformation $U = F(X)$. If $F(X)$ is the distribution function of X, the random variable U is uniformly distributed between 0 and 1.

Given n observations $X_{(1)}, \ldots, X_{(n)}$, PROC UNIVARIATE computes the values $U_{(i)} = F\left(X_{(i)}\right)$ by applying the transformation, as follows.

When you specify the NORMAL option in the PROC UNIVARIATE statement or use the HISTOGRAM statement to fit a parametric distribution, PROC UNIVARIATE provides a series of goodness-of-fit tests that are based on the empirical distribution function (EDF):

- Kolmogorov-Smirnov
- Anderson-Darling
- Cramer-von Mises

These tests are based on various measures of the discrepancy between the empirical distribution function $F_n(x)$ and the proposed cumulative distribution function $F(x)$.

Once the EDF test statistics are computed, the associated *p*-values are calculated. PROC UNIVARIATE uses internal tables of probability levels that are similar to those given by D'Agostino and Stephens (1986). If the value lies between two probability levels, then linear interpolation is used to estimate the probability value.

Note: PROC UNIVARIATE does not support some of the EDF tests when you use the HISTOGRAM statement and you estimate the parameters of the specified distribution. See Table 41.7 on page 1400 for more information. △

Kolmogorov *D* Statistic

The Kolmogorov-Smirnov statistic (D) is defined as

$$D = \sup_x \left| F_n(x) - F_{(x)} \right|$$

The Kolmogorov-Smirnov statistic belongs to the supremum class of EDF statistics. This class of statistics is based on the largest vertical difference between $F(x)$ and $F_n(x)$.

The Kolmogorov-Smirnov statistic is computed as the maximum of D^+ and D^-. D^+ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function. D^- is the largest vertical distance when the EDF is less than the distribution function.

$$D^+ = \max_i \left(\frac{i}{n} - U_{(i)} \right)$$
$$D^- = \max_i \left(U_{(i)} - \frac{i-1}{n} \right)$$
$$D = \max \left(D^+, D^- \right)$$

PROC UNIVARIATE uses a modified Kolmogorov *D* statistic to test the data against a normal distribution with mean and variance equal to the sample mean and variance.

Anderson-Darling Statistic

The Anderson-Darling statistic and the Cramer-von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference $\left(F_n(x) - F(x)\right)^2$. Quadratic statistics have the following general form:

$$Q = n \int_{-\infty}^{+\infty} \left(F_n(x) - F(x)\right)^2 \psi(x)\, dF(x)$$

The function $\psi(x)$ weights the squared difference $\left(F_n(x) - F(x)\right)^2$.

The Anderson-Darling statistic (A^2) is defined as

$$A^2 = n \int_{-\infty}^{+\infty} \left(F_n(x) - F(x)\right)^2 \left[F(x)\left(1 - F(x)\right)\right]^{-1} dF(x)$$

where the weight function is $\psi(x) = \left[F(x)\left(1 - F(x)\right)\right]^{-1}$.

The Anderson-Darling statistic is computed as

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1) \left(\log U_{(i)} + \log\left(1 - U_{(n+1-i)}\right) \right) \right]$$

Cramer-von Mises Statistic

The Cramer-von Mises statistic (W^2) is defined as

$$W^2 = n \int_{-\infty}^{+\infty} \left(F_n(x) - F(x)\right)^2 dF(x)$$

where the weight function is $\psi(x) = 1$.

The Cramer-von Mises statistic is computed as

$$W^2 = \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

Probability Values of EDF Tests

Once the EDF test statistics are computed, PROC UNIVARIATE computes the associated probability values.

The probability value depends upon the parameters that are known and the parameters that PROC UNIVARIATE estimates for the fitted distribution. Table 41.7 on page 1400 summarizes different combinations of estimated parameters for which EDF tests are available.

Note: PROC UNIVARIATE assumes that the threshold (THETA=) parameter for the beta, exponential, gamma, lognormal, and Weibull distributions is known. If you omit its value, PROC UNIVARIATE assumes that it is zero and that it is known. Likewise, PROC UNIVARIATE assumes that the SIGMA= parameter, which determines the upper threshold (SIGMA) for the beta distribution, is known. If you omit its value, PROC UNIVARIATE assumes that the value is one. These parameters are not listed in Table 41.7 on page 1400 because they are assumed to be known in all cases, and they do not affect which EDF statistics PROC UNIVARIATE computes. △

Table 41.7 Availability of EDF Tests

Distribution	Parameters	EDF
Beta	α and β unknown	none
	α known, β unknown	none
	α unknown, β known	none
	α and β known	all
Exponential	σ unknown	all
	σ known	all
Gamma	α and σ unknown	none
	α known, σ unknown	none
	α unknown, σ known	none
	α and σ known	all
Lognormal	ζ and σ unknown	all
	ζ known, σ unknown	A^2 and W^2
	ζ unknown, σ known	A^2 and W^2
	ζ and σ known	all

Distribution	Parameters	EDF
Normal	μ and σ unknown	all
	μ known, σ unknown	A^2 and W^2
	μ unknown, σ known	A^2 and W^2
	μ and σ known	all
Weibull	c and σ unknown	A^2 and W^2
	c known, σ unknown	A^2 and W^2
	c unknown, σ known	A^2 and W^2
	c and σ known	all

Robust Estimators

A statistical method is robust if the method is insensitive to slight departures from the assumptions that justify the method. PROC UNIVARIATE provides several methods for robust estimation of location and scale.

Winsorized Means

When outliers are present in the data, the Winsorized mean is a robust estimator of the location that is relatively insensitive to the outlying values. The *k*-times Winsorized mean is calculated as

$$\overline{x}_{wk} = \frac{1}{n}\left((k+1)\, x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)\, x_{(n-k)} \right)$$

The Winsorized mean is computed after the k smallest observations are replaced by the $(k+1)$ smallest observation, and the k largest observations are replaced by the $(k+1)$ largest observation.

For a symmetric distribution, the symmetrically Winsorized mean is an unbiased estimate of the population mean. But the Winsorized mean does not have a normal distribution even if the data are from a normal population.

The Winsorized sum of squared deviations is defined as

$$\begin{aligned} s^2_{wk} = {} & (k+1)\left(x_{(k+1)} - \overline{x}_{wk}\right)^2 + \sum_{i=k+2}^{n-k-1} \left(x_{(i)} - \overline{x}_{wk}\right)^2 \\ & + (k+1)\left(x_{(n-k)} - \overline{x}_{wk}\right)^2 \end{aligned}$$

A Winsorized t test is given by

$$t_{wk} = \frac{(\overline{x}_{wk} - \mu_0)}{STDERR\,(\overline{x}_{wk})}$$

where the standard error of the Winsorized mean is

$$STDERR(\overline{x}_{wk}) = \frac{n-1}{n-2k-1}\frac{s_{wk}}{\sqrt{n(n-1)}}$$

When the data are from a symmetric distribution, the distribution of the Winsorized t statistic t_{wk} is approximated by a Student's t distribution with $n-2k-1$ degrees of freedom (Tukey and McLaughlin 1963, Dixon and Tukey 1968).

A $100(1-\alpha)$ percent confidence interval for the Winsorized mean has upper and lower limits

$$\overline{x}_{wk} \pm t_{(1-\alpha/2)} STDERR(\overline{x}_{wk})$$

and the $(1-\alpha/2)$ critical value of the Student's t statistics has $n-2k-1$ degrees of freedom.

Trimmed Means

When outliers are present in the data, the trimmed mean is a robust estimator of the location that is relatively insensitive to the outlying values. The k-times trimmed mean is calculated as

$$\overline{x}_{tk} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}$$

The trimmed mean is computed after the k smallest and k largest observations are deleted from the sample. In other words, the observations are trimmed at each end.

For a symmetric distribution, the symmetrically trimmed mean is an unbiased estimate of the population mean. But the trimmed mean does not have a normal distribution even if the data are from a normal population.

A robust estimate of the variance of the trimmed mean t_{tk} can be based on the Winsorized sum of squared deviations (Tukey and McLaughlin 1963). The resulting trimmed t test is given by

$$t_{tk} = \frac{(\overline{x}_{tk} - \mu_0)}{STDERR(\overline{x}_{tk})}$$

where the standard error of the trimmed mean is

$$STDERR(\overline{x}_{tk}) = \frac{s_{wk}}{\sqrt{(n-2k)(n-2k-1)}}$$

and s_{wk} is the square root of the Winsorized sum of squared deviations

When the data are from a symmetric distribution, the distribution of the trimmed t statistic t_{tk} is approximated by a Student's t distribution with $n-2k-1$ degrees of freedom (Tukey and McLaughlin 1963, Dixon and Tukey 1968).

A $100(1-\alpha)$ percent confidence interval for the trimmed mean has upper and lower limits

$$\overline{x}_{tk} \pm t_{(1-\alpha/2)} STDERR\left(\overline{x}_{tk}\right)$$

and the $(1-\alpha/2)$ critical value of the Student's *t* statistics has $n-2k-1$ degrees of freedom.

Robust Measures of Scale

The sample standard deviation is a commonly used estimator of the population scale. However, it is sensitive to outliers and may not remain bounded when a single data point is replaced by an arbitrary number. With robust scale estimators, the estimates remain bounded even when a portion of the data points are replaced by arbitrary numbers.

PROC UNIVARIATE computes robust measures of scale that include statistics of interquartile range, Gini's mean difference *G*, *MAD*, Q_n, and S_n, with their corresponding estimates of σ.

The interquartile range is a simple robust scale estimator, which is the difference between the upper and lower quartiles. For a normal population, the standard deviation σ can be estimated by dividing the interquartile range by 1.34898.

Gini's mean difference is also a robust estimator of the standard deviation σ. For a normal population, Gini's mean difference has expected value $2\sigma/\sqrt{\pi}$. Thus, multiplying Gini's mean difference by $\sqrt{\pi}/2$ yields a robust estimator of the standard deviation when the data are from a normal sample. The constructed estimator has high efficiency for the normal distribution relative to the usual sample standard deviation. It is also less sensitive to the presence of outliers than the sample standard deviation.

Gini's mean difference is computed as

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j|$$

If the observations are from a normal distribution, then $\sqrt{\pi}\, G/2$ is an unbiased estimator of the standard deviation σ.

A very robust scale estimator is the *MAD*, the median absolute deviation about the median (Hampel, 1974.)

$$MAD = \mathrm{med}_i\left(|x_i - \mathrm{med}_j\left(x_j\right)|\right)$$

where the inner median, $\mathrm{med}_j(x_j)$, is the median of the n observations and the outer median, med_i, is the median of the n absolute values of the deviations about the median.

For a normal distribution, 1.4826·*MAD* can be used to estimate the standard deviation σ.

The *MAD* statistic has low efficiency for normal distributions, and it may not be appropriate for symmetric distributions. Rousseeuw and Croux (1993) proposed two new statistics as alternatives to the *MAD* statistic.

The first statistic is

$$S_n = 1.1926\, \mathrm{med}_i(\mathrm{med}_j(|x_i - x_j|))$$

where the outer median, med_i, is the median of the n medians of $(|x_i - x_j|);\ \ j = 1, 2, \ldots, n$.

To reduce the small-sample bias, $c_{sn} S_n$ is used to estimate the standard deviation σ, where c_{sn} is a the correction factor (Croux and Rousseeuw, 1992.)

The second statistic is

$$Q_n = 2.219 \left\{ |x_i - x_j|;\ \ i < j \right\}_{(k)}$$

where $k = \binom{h}{2}$, $h = [n/2] + 1$, and $[n/2]$ is the integer part of $n/2$. That is, Q_n is 2.2219 times the kth order statistic of the $\binom{n}{2}$ distances between data points.

The bias-corrected statistic, $c_{qn}\, Q_n$, is used to estimate the standard deviation σ, where c_{qn} is a correction factor.

Calculating Percentiles

The UNIVARIATE procedure automatically computes the minimum, 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 99th, and maximum percentiles. You use the PCTLDEF= option in the PROC UNIVARIATE statement to specify one of five methods to compute quantile statistics. See for more information.

To compute the quantile that each observation falls in, use PROC RANK with the GROUP= option. To calculate percentiles other than the default percentiles, use PCTLPTS= and PCTLPRE= in the OUTPUT statement.

Confidence Limits for Quantiles

The CIPCTLDF option and CIPCTLNORMAL option compute confidence limits for quantiles using methods described in Hahn and Meeker (1991).

When $0.0 < p < 0.5$, the two-sided $100\,(1 - \alpha)$ percent confidence interval for quantiles that are based on normal data has lower and upper limits

$$\overline{x} - g'_{(\alpha/2;1-p,n)} s, \overline{x} - g'_{(1-\alpha/2;1-p,n)} s$$

where p is the percentile $100 \times p$.

When $0.5 \leq p < 1.0$, the lower and upper limits are

$$\overline{x} + g'_{(\alpha/2;p,n)} s,\ \overline{x} + g'_{(1-\alpha/2;p,n)} s$$

A one-sided $100\,(1 - \alpha)$ percent confidence limit is computed by replacing $\alpha/2$ with α. The factor $g'_{(\gamma,p,n)}$ is described in Owen and Hua (1977) and Odeh and Owen (1980).

The two-sided distribution-free $100\,(1 - \alpha)\%$ confidence interval for quantiles from a sample of size n is

$$x_{(l)}, x_{(u)}$$

where $x_{(j)}$ is *j*th order statistic. The lower rank l and upper rank u are integers that are symmetric or nearly symmetric around $i = [np] + 1$, where $[np]$ is the integral part of np.

The l and u are chosen so that the order statistics $x_{(l)}$ and $x_{(u)}$

- are approximately symmetric about $x_{((n+1)p)}$
- are as close to $x_{((n+1)p)}$ as possible
- satisfy the coverage probability requirement.

$$Q_b(u-1;n,p) - Q_b(l-1;n,p) \geq 1-\alpha$$

where Q_b is the cumulative binomial probability, $0 < l < u \leq n$, and $0 < p < 1$.

The coverage probability is sometimes less that $1-\alpha$. This can occur in the tails of the distribution when the sample size is small. To avoid this problem, you can specify the option TYPE=ASYMMETRIC, which causes PROC UNIVARIATE to use asymmetric values of l and u. However, PROC UNIVARIATE first attempts to compute confidence limits that satisfy all three conditions. If the last condition is not satisfied, then the first condition is relaxed. Thus, some of the confidence limits may be symmetric while others, especially in the extremes, are not.

A one-sided distribution-free lower $100(1-\alpha)$ percent confidence limit is computed as $x_{(l)}$ when l is the largest integer that satisfies the inequality

$$1 - Q_b(l-1;n,p) \geq 1-\alpha$$

where $0 < l \leq n$, and $0 < p < 1$. Likewise, a one-sided distribution-free upper $100(1-\alpha)$% confidence limit is computed as $x_{(u)}$ when u is the smallest integer that satisfies the inequality

$$Q_b(u-1;n,p) \geq 1-\alpha$$

where $0 < u \leq n$, and $0 < p < 1$.

Weighted Quantiles

When you use the WEIGHT statement the percentiles are computed as follows. Let x_i be the ith ordered nonmissing value, $x_1 \leq x_2 \leq \ldots \leq x_n$. Then, for a given value of p between 0 and 1, the pth weighted quantile (or 100 pth weighted percentile), y, is computed from the empirical distribution function with averaging

$$y = \begin{cases} \frac{1}{2}(x_i + x_{i+1}) & \text{if } \sum_{j=1}^{i} w_j = pW \\ x_{i+1} & \text{if } \sum_{j=1}^{i} w_j < pW < \sum_{j=1}^{i+1} w_j \end{cases}$$

where w_j is the weight associated with x_i, $W = \sum_{i=1}^{n} w_i$ is the sum of the weights and w_i is the weight for ith observation.

When the observations have identical weights, the weighted percentiles are the same as the unweighted percentiles with PCTLDEF=5.

Calculating the Mode

The mode is the value that occurs most often in the data. PROC UNIVARIATE counts repetitions of the actual values or, if you specify the ROUND= option, the rounded values. If a tie occurs for the most frequent value, the procedure reports the lowest value. To list all possible modes, use the MODES option in the PROC UNIVARIATE statement. When no repetitions occur in the data (as with truly continuous data), the procedure does not report the mode.

The WEIGHT statement has no effect on the mode.

Formulas for Fitted Continuous Distributions

The following sections provide information about the families of parametric distributions that you can fit with the HISTOGRAM statement. Properties of the parametric curves are discussed by Johnson, et al. (1994).

Beta Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} h \times 100\% & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta+\sigma \end{cases}$$

where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

$\theta =$ lower threshold parameter (lower endpoint parameter)
$\sigma =$ scale parameter $(\sigma > 0)$
$\alpha =$ shape parameter $(\alpha > 0)$
$\beta =$ shape parameter $(\beta > 0)$
$h =$ width of histogram interval

This notation is consistent with that of other distributions that you can fit with the HISTOGRAM statement. However, many texts, including Johnson, et al. (1994), write the beta density function as:

$$p(x) = \begin{cases} \frac{(x-a)^{p-1}(b-x)^{q-1}}{\mathrm{B}(p,q)(b-a)^{p+q-1}} & \text{for } a < x < b \\ 0 & \text{for } x \leq a \text{ or } x \geq b \end{cases}$$

The two notations are related as follows:

$$\sigma = b - a$$
$$\theta = a$$
$$\alpha = p$$
$$\beta = q$$

The range of the beta distribution is bounded below by a threshold parameter $\theta = a$ and above by $\theta + \sigma = b$. If you specify a fitted beta curve using the BETA option, θ must be less than the minimum data value, and $\theta + \sigma$ must be greater than the maximum data value. You can specify θ and σ with the THETA= and SIGMA= *value* in parentheses after the keyword BETA. By default, $\sigma = 1$ and $\theta = 0$. If you specify THETA=EST and SIGMA=EST, maximum likelihood estimates are computed for θ and σ.

Note: However, three- and four-parameter maximum likelihood estimation may not always converge. △

In addition, you can specify α and β with the ALPHA= and BETA= *beta-options*, respectively. By default, the procedure calculates maximum likelihood estimates for α and β. For example, to fit a beta density curve to a set of data bounded below by 32 and above by 212 with maximum likelihood estimates for α and β, use the following statement:

```
histogram length / beta(theta=32 sigma=180);
```

The beta distributions are also referred to as Pearson Type I or II distributions. These include the *power-function* distribution ($\beta = 1$), the *arc-sine* distribution ($\alpha = \beta = \frac{1}{2}$), and the generalized *arc-sine* distributions ($\alpha + \beta = 1, \beta \neq \frac{1}{2}$). You can use the DATA step function BETAINV to compute beta quantiles and the DATA step function PROBBETA to compute beta probabilities.

Exponential Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{h \times 100\%}{\sigma} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for} x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where

$$\theta = \text{threshold parameter}$$
$$\sigma = \text{scale parameter } (\sigma > 0)$$
$$h = \text{width of histogram interval}$$

The threshold parameter θ must be less than or equal to the minimum data value. You can specify θ with the THRESHOLD= *exponential-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. In addition, you can specify σ with the SCALE= *exponential-option*. By default, the procedure calculates a maximum likelihood estimate for σ. Note that some authors define the scale parameter as $\frac{1}{\sigma}$.

The exponential distribution is a special case of both the gamma distribution (with $\alpha = 1$ and the Weibull distribution (with $c = 1$). A related distribution is the *extreme value* distribution. If $Y = \exp(-\mathrm{X})$ has an exponential distribution, then X has an extreme value distribution.

Gamma Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{h \times 100\%}{\Gamma(\alpha)\sigma} \left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for} x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

$$\begin{aligned} \theta &= \text{threshold parameter} \\ \sigma &= \text{scale parameter } (\sigma > 0) \\ \alpha &= \text{shape parameter } (\alpha > 0) \\ h &= \text{width of histogram interval} \end{aligned}$$

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THRESHOLD= *gamma-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. In addition, you can specify σ and α with the SCALE= and ALPHA= *gamma-options*. By default, the procedure calculates maximum likelihood estimates for σ and α.

The gamma distributions are also referred to as Pearson Type III distributions, and they include the chi-square, exponential, and Erlang distributions. The probability density function for the chi-square distribution is

$$p(x) = \begin{cases} \frac{1}{2\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{x}{2}\right)^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right) & \text{for} x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

Notice that this is a gamma distribution with $\alpha = \frac{\nu}{2}$, and $\theta = 0$. The exponential distribution is a gamma distribution with $\alpha = 1$, and the Erlang distribution is a gamma distribution with α being a positive integer. A related distribution is the Rayleigh distribution. If $R = \frac{\max(X_1,\ldots,X_n)}{\min(X_1,\ldots,X_n)}$ where the X_i's are independent χ^2_ν variables, then $\log R$ is distributed with a χ_ν distribution having a probability density function of

$$p(x) = \begin{cases} \left[2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)\right]^{-1} x^{\nu-1} \exp\left(-\frac{x^2}{2}\right) & \text{for} x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

If $\nu = 2$, the preceding distribution is referred to as the Rayleigh distribution. You can use the DATA step function GAMINV to compute gamma quantiles and the DATA step function PROBGAM to compute gamma probabilities.

Lognormal Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{h \times 100\%}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

$$\theta = \text{threshold parameter}$$
$$\zeta = \text{scale parameter}\ (-\infty < \zeta < \infty)$$
$$\sigma = \text{shape parameter}\ (\sigma > 0)$$
$$h = \text{width of histogram interval}$$

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THRESHOLD= *lognormal-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify ζ and σ with the SCALE= and SHAPE= *lognormal-options*, respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.

Note: σ denotes the shape parameter of the lognormal distribution, whereas σ denotes the scale parameter of the beta, exponential, gamma, normal, and Weibull distributions. The use of σ to denote the lognormal shape parameter is based on the fact that $\frac{1}{\sigma}(\log(X-\theta)-\zeta)$ has a standard normal distribution if X is lognormally distributed. △

Normal Distribution

The fitted density function is

$$p(x) = \frac{h \times 100\%}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{for } -\infty < x < \infty$$

where

$$\mu = \text{mean}$$
$$\sigma = \text{standard deviation}\ (\sigma > 0)$$
$$h = \text{width of histogram interval}$$

You can specify μ and σ with the MU= and SIGMA= *normal-options*, respectively. By default, the procedure estimates μ with the sample mean and σ with the sample standard deviation. You can use the DATA step function PROBIT to compute normal quantiles and the DATA step function PROBNORM to compute probabilities.

Weibull Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{ch \times 100\%}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right) & \text{for} x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

$$\theta = \text{threshold parameter}$$
$$\sigma = \text{scale parameter } (\sigma > 0)$$
$$c = \text{shape parameter } (\alpha > 0)$$
$$h = \text{width of histogram interval}$$

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THRESHOLD= *Weibull-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify σ and c with the SCALE= and SHAPE= *Weibull-options*, respectively. By default, the procedure calculates maximum likelihood estimates for σ and c.

The exponential distribution is a special case of the Weibull distribution where $c = 1$.

Kernel Density Estimates

You can use the KERNEL option to superimpose kernel density estimates on histograms. Smoothing the data distribution with a kernel density estimate can be more effective than using a histogram to visualize features that might be obscured by the choice of histogram bins or sampling variation. For example, a kernel density estimate can also be more effective when the data distribution is multimodal. The general form of the kernel density estimator is

$$\hat{f}_\lambda(x) = \frac{1}{n_\lambda} \sum_{i=1}^{n} K_0\left(\frac{x - x_i}{\lambda}\right)$$

where $K_0(.)$ is a kernel function, λ is the bandwidth, n is the sample size, and x_i is the i^{th} observation.

The KERNEL option provides three kernel functions (K_0): normal, quadratic, and triangular. You can specify the function with the K=*kernel-function* in parentheses after the KERNEL option. Values for the K= option are NORMAL, QUADRATIC, and TRIANGULAR (with aliases of N, Q, and T, respectively). By default, a normal kernel is used. The formulas for the kernel functions are

Normal $\quad K_0(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right)$ for $-\infty < t < \infty$

Quadratic $\quad K_0(t) = \frac{3}{4}\left(1 - t^2\right)$ for $|t| \leq 1$

Triangular $\quad K_0(t) = 1 - |t|$ for $|t| \leq 1$

The value of λ, referred to as the bandwidth parameter, determines the degree of smoothness in the estimated density function. You specify λ indirectly by specifying a standardized bandwidth c with the C=*kernel-option*. If Q is the interquartile range, and n is the sample size, then c is related to λ by the formula

$$\lambda = cQn^{-\frac{1}{5}}$$

For a specific kernel function, the discrepancy between the density estimator $\hat{f}_\lambda(x)$ and the true density $f(x)$ is measured by the mean integrated square error (MISE):

$$\mathrm{MISE}(\lambda) = \int_x \left\{ E\left(\hat{f}_\lambda(x)\right) - f(x) \right\}^2 dx + \int_x var\left(\hat{f}_\lambda(x)\right) dx$$

The MISE is the sum of the integrated squared bias and the variance. An approximate mean integrated square error (AMISE) is

$$\mathrm{AMISE}(\lambda) = \frac{1}{4}\lambda^4 \left(\int_{\mathrm{t}} \mathrm{t}^2 K(t)\, dt \right)^2 \int_x \left(f''(x)\right)^2 dx + \frac{1}{\mathrm{n}\lambda} \int_{\mathrm{t}} K(t)^2\, dt$$

A bandwidth that minimizes AMISE can be derived by treating $f(x)$ as the normal density having parameters μ and σ estimated by the sample mean and standard deviation. If you do not specify a bandwidth parameter or if you specify C=MISE, the bandwidth that minimizes AMISE is used. The value of AMISE can be used to compare different density estimates. For each estimate, the bandwidth parameter c, the kernel function type, and the value of AMISE are reported in the SAS log.

Theoretical Distributions for Quantile-Quantile and Probability Plots

You can use the PROBPLOT and QQPLOT statements to request probability and Q-Q plots that are based on the theoretical distributions that are summarized in the following table:

Table 41.8 Distributions and Parameters

Distribution	Density Function $p(x)$	Range	Parameters		
			Location	Scale	Shape
Beta	$\frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{\beta(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}$	$\theta < x < \theta+\sigma$	θ	σ	α, β
Exponential	$\frac{1}{\sigma}\exp\left(-\frac{x-\theta}{\sigma}\right)$	$x \geq \theta$	θ	σ	
Gamma	$\frac{1}{\sigma\Gamma(\alpha)}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{x-\theta}{\sigma}\right)$	$x > \theta$	θ	σ	α
Lognormal (3-parameter)	$\frac{1}{\sigma\sqrt{2\pi}(x-\theta)}\exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right)$	$x > \theta$	θ	ζ	σ
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$all\ x$	μ	σ	
Weibull (3-parameter)	$\frac{c}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right)$	$x > \theta$	θ	σ	c
Weibull2 (2-parameter)	$\frac{c}{\sigma}\left(\frac{x-\theta_0}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta_0}{\sigma}\right)^c\right)$	$x > \theta_0$	θ_0 (known)	σ	c

You can request these distributions with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, WEIBULL, and WEIBULL2 options, respectively. If you

omit a distribution option, the PROBPLOT statement creates a normal probability plot and the QQPLOT statement creates a normal Q-Q plot.

The following sections provide the details for constructing Q-Q plots that are based on these distributions. Probability plots are constructed similarly except that the horizontal axis is scaled in percentile units.

Beta Distribution

To create a plot that is based on the beta distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the i^{th} ordered observation against the quantile $\mathrm{B}_{\alpha\beta}^{-1}\left(\frac{i-0.375}{n+0.25}\right)$ where $\mathrm{B}_{\alpha\beta}^{-1}(.)$ is the inverse normalized incomplete beta function, n is the number of nonmissing observations, and α and β are the shape parameters of the beta distribution.

The point pattern on the plot for ALPHA=α and BETA=β tends to be linear with intercept θ and slope σ if the data are beta distributed with the specific density function

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and θ is the lower threshold parameter, σ is the scale parameter $(\sigma > 0)$, α the first shape parameter $(\alpha > 0)$ and β is the second shape parameter $(\beta > 0)$.

Exponential Distribution

To create a plot that is based on the exponential distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the i^{th} ordered observation against the quantile $-\log\left(1-\frac{i-0.375}{n+0.25}\right)$ where n is the number of nonmissing observations.

The point pattern on the plot tends to be linear with intercept θ and slope σ if the data are exponentially distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma}\exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for} x \ge \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where θ is a threshold parameter and σ is a positive scale parameter.

Gamma Distribution

To create a plot that is based on the gamma distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the i^{th} ordered observation against the quantile $G_{\alpha}^{-1}\left(\frac{i-0.375}{n+0.25}\right)$ where G_{α}^{-1} is the inverse normalized incomplete gamma function, n is the number of nonmissing observations, and α is the shape parameter of the gamma distribution.

The point pattern on the plot tends to be linear with intercept θ and slope σ if the data are gamma distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma\Gamma(\alpha)}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for} x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where θ is the threshold parameter, σ is the scale parameter $(\sigma > 0)$, and α is the shape parameter $(\alpha > 0)$.

Lognormal Distribution

To create a plot that is based on the lognormal distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the i^{th} ordered observation against the quantile $\exp\left(\sigma\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)\right)$ where $\Phi^{-1}(.)$ is the inverse standard cumulative normal distribution, n is the number of nonmissing observations, and σ is the shape parameter of the lognormal distribution.

The point pattern on the plot for SIGMA=σ tends to be linear with intercept θ and slope $\exp(\zeta)$ if the data are lognormally distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for} x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where θ is the threshold parameter, ζ is the scale parameter, and σ is the shape parameter $(\sigma > 0)$.

Normal Distribution

To create a plot that is based on the normal distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the i^{th} ordered observation against the quantile $\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$ where $\Phi^{-1}(.)$ is the inverse cumulative standard normal distribution and n is the number of nonmissing observations.

The point pattern on the plot tends to be linear with intercept μ and slope σ if the data are normally distributed with the specific density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for all } x$$

where μ is the mean and σ is the standard deviation $(\sigma > 0)$.

Three-Parameter Weibull Distribution

To create a plot that is based on a three-parameter Weibull distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the i^{th} ordered observation against the quantile $\left(-\log\left(1-\frac{i-0.375}{n+0.25}\right)\right)^{\frac{1}{c}}$ where n is the number of nonmissing observations, and α and c are the Weibull distribution shape parameters.

The point pattern on the plot for C=c tends to be linear with intercept θ and slope σ if the data are Weibull distributed with the specific density function

$$p(x) = \begin{cases} \frac{c}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right) & \text{for} x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where θ is the threshold parameter, σ is the scale parameter $(\sigma > 0)$, and c is the shape parameter $(c > 0)$.

Two-Parameter Weibull Distribution

To create a plot that is based on a two-parameter Weibull distribution, PROC UNIVARIATE orders the observations from smallest to largest, and plots the log of the shifted i^{th} ordered observation $x_{(i)}$, denoted by $\log\left(x_{(i)} - \theta_0\right)$, against the quantile $\left(-\log\left(1 - \frac{i-0.375}{n+0.25}\right)\right)$ where n is the number of nonmissing observations.

Unlike the three-parameter Weibull quantile, the preceding expression is free of distribution parameters. This is why the C= shape parameter is not required in the WEIBULL2 option.

The point pattern on the plot for THETA=θ_0 tends to be linear with intercept $\log(\sigma)$ and slope $\frac{1}{c}$ if the data are Weibull distributed with the specific density function

$$p(x) = \begin{cases} \frac{c}{\sigma}\left(\frac{x-\theta_0}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta_0}{\sigma}\right)^{c}\right) & \text{for} x > \theta_0 \\ 0 & \text{for } x \leq \theta_0 \end{cases}$$

where θ_0 is the known lower threshold, σ is the scale parameter $(\sigma > 0)$, and c is the shape parameter $(c > 0)$.

Shape Parameters

Some distribution options in the PROBPLOT and QQPLOT statements require that you specify one or two shape parameters in parentheses after the distribution keyword. These are summarized in the following table:

Table 41.9 Shape Parameter Options

Distribution Keyword	Required Shape Parameter Option	Range
BETA	ALPHA=α, BETA=β	$\alpha > 0,\ \beta > 0$
EXPONENTIAL	None	
GAMMA	ALPHA=α	$\alpha > 0$
LOGNORMAL	SIGMA=σ	$\sigma > 0$
NORMAL	None	
WEIBULL	C=c	$c > 0$
WEIBULL2	None	

You can visually estimate the value of a shape parameter by specifying a list of values for the shape parameter option. PROC UNIVARIATE produces a separate plot for each value. Then you can use the value of the shape parameter that produces the most nearly linear point pattern. Alternatively, you can request that PROC UNIVARIATE use an estimated shape parameter to create the plot.

Note: For Q-Q plots that are requested with the WEIBULL2 option, you can estimate the shape parameter c from a linear pattern by using the fact that the slope of the pattern is $\frac{1}{c}$. △

Location and Scale Parameters

When you use the PROBPLOT statement to specify or estimate the location and scale parameters for a distribution, a diagonal distribution reference line appears on the probability plot. (An exception is the two-parameter Weibull distribution, where the line appears when you specify or estimate the scale and shape parameters.) Agreement between this line and the point pattern indicates that the distribution with these parameters is a good fit.

Note: Close visual agreement may not necessarily mean that the distribution is a good fit based on the criteria that are used by formal goodness-of-fit tests. △

When the point pattern on a Q-Q plot is linear, its intercept and slope provide estimates of the location and scale parameters. (An exception to this rule is the two-parameter Weibull distribution, for which the intercept and slope are related to the scale and shape parameters.) When you use the QQPLOT statement to specify or estimate the slope and intercept of the line, a diagonal distribution reference line appears on the Q-Q plot. This line allows you to check the linearity of the point pattern.

The following table shows which parameters to specify to determine the intercept and slope of the line:

Table 41.10 Intercept and Slope of Distribution Reference Line

	Parameters			Linear Pattern	
Distribution	Location	Scale	Shape	Intercept	Slope
BETA	θ	σ	α, β	θ	σ
EXPONENTIAL	θ	σ		θ	σ
GAMMA	θ	σ	α	θ	σ
LOGNORMAL	θ	ζ	σ	θ	$\exp(\zeta)$
NORMAL	μ	σ		μ	σ
WEIBULL (3-parameter)	θ	σ	c	θ	σ
WEIBULL2 (2-parameter)	θ_0 (known)	σ	c	$\log(\sigma)$	$\frac{1}{c}$

For the LOGNORMAL and WEIBULL2 options, you can specify the slope directly with the SLOPE= option. That is, for the LOGNORMAL option, when you specify THETA=θ_0 and SLOPE=$\exp(\zeta_0)$, PROC UNIVARIATE displays the same line as that which is specified by THETA=θ_0 and ZETA=ζ_0. For the WEIBULL2 option, when you specify SIGMA=σ_0 and SLOPE=$\frac{1}{c_0}$, PROC UNIVARIATE displays the same line when you specify SIGMA=σ_0 and C=c_0. Alternatively, you can request to use the estimated values of the parameters to determine the reference line.

Results

By default, PROC UNIVARIATE produces tables of moments, basic statistical measures, tests for location, quantiles, and extreme observations. You must specify options in the PROC UNIVARIATE statement to produce other statistics and tables.

The CIBASIC option produces the table of the basic confidence measures that includes the confidence limits for the mean, standard deviation, and variance. The CIPCTLDF option and CIPCTLNORMAL option produce tables of confidence limits for the quantiles. The LOCCOUNT option produces the table that shows the number of values greater than, equal to, and less than the value of MU0=. The FREQ option produces the table of frequencies counts. The NEXTRVAL= option produces the table with the frequencies of the extreme values. The NORMAL option produces the table with the tests for normality. The TRIMMED=, WINSORIZED=, and ROBUSTCALE options produce tables with robust estimators.

The table of trimmed or Winsorized means includes the percentage and the number of observations that are trimmed or Winsorized at each end, the mean and standard error, confidence limits, and the Student's t test. The table with robust measures of scale includes interquartile range, Gini's mean difference G, MAD, Q_n, and S_n, with their corresponding estimates of σ.

Missing Values

PROC UNIVARIATE excludes missing values for the analysis variable before calculating statistics. Each analysis variable is treated individually; a missing value for an observation in one variable does not affect the calculations for other variables. The statements handle missing values as follows:

- If a BY or an ID variable value is missing, PROC UNIVARIATE treats it like any other BY or ID variable value. The missing values form a separate BY group.
- If the FREQ variable value is missing or nonpositive, PROC UNIVARIATE excludes the observation from the analysis.
- If the WEIGHT variable value is missing, PROC UNIVARIATE excludes the observation from the analysis.

PROC UNIVARIATE tabulates the number of the missing values and reports this information in the procedure output. Before the number of missing values is tabulated, PROC UNIVARIATE excludes observations when

- you use the FREQ statement and the frequencies are nonpositive
- you use the WEIGHT statement and the weights are missing or nonpositive (you must specify the EXCLNPWGT option).

Histograms

If you request a fitted parametric distribution with a HISTOGRAM statement, PROC UNIVARIATE creates a report that summarizes the fit in addition to the graphical display. The report includes information about

- parameters for the fitted curve, estimated mean, and estimated standard deviation
- EDF goodness-of-fit tests
- histogram intervals
- quantiles.

Histogram Intervals

If you specify the MIDPERCENTS suboption in parentheses after a density estimate option, PROC UNIVARIATE includes a table that lists the interval midpoints along with the observed and estimated percentages of the observations that lie in the interval.

The estimated percentages are based on the fitted distribution. You can also specify the MIDPERCENTS suboption to request a table of interval midpoints with the observed percentage of observations that lie in the interval.

Quantiles

By default, PROC UNIVARIATE displays a table that lists observed and estimated quantiles for the 1, 5, 10, 25, 50, 75, 90, 95, and 99 percent of a fitted parametric distribution. You can use the PERCENTS= suboption to request that the quantiles for specfic percentiles appear in the table.

Output Data Set

PROC UNIVARIATE can create one or more output SAS data sets. When you specify an OUTPUT statement and no BY statement, PROC UNIVARIATE creates an output data set that contains one observation. If you use a BY statement, the corresponding output data set contains an observation with statistics for each BY group. The procedure does not print the output data set. Use PROC PRINT, PROC REPORT, or another SAS reporting tool to print the output data set.

The output data set includes

- BY statement variables
- variables that contain statistics
- variables that contain percentiles.

The BY variables indicate which BY group each observation summarizes. When you omit a BY statement, the procedure computes statistics and percentiles by using all the observations in the input data set. When you use a BY statement, the procedure computes statistics and percentiles by using the observations within each BY group.

OUTHISTOGRAM= Data Set

You can create a OUTHISTOGRAM= data in the HISTOGRAM statement that contains information about histogram intervals. Because you can specify multiple HISTOGRAM statements with the UNIVARIATE procedure, you can create multiple OUTHISTOGRAM= data sets.

The data set contains a group of observations for each variable that the HISTOGRAM statement plots. The group contains an observation for each interval of the histogram, beginning with the leftmost interval that contains a value of the variable and ending with the rightmost interval that contains a value of the variable. These intervals will not necessarily coincide with the intervals displayed in the histogram since the histogram may be padded with empty intervals at either end. If you superimpose one or more fitted curves on the histogram, the OUTHISTOGRAM= data set contains multiple groups of observations for each variable (one group for each curve). If you use a BY statement, the OUTHISTOGRAM= data set contains groups of observations for each BY group. ID variables are not saved in the OUTHISTOGRAM= data set.

The variables in OUTHISTOGRAM= data set are

CURVE	name of fitted distribution (if requested in HISTOGRAM statement)
EXPPCT	estimated percent of population in histogram interval determined from optional fitted distribution
MIDPT	midpoint of fitted distribution

OBSPCT	percent of variable values in histogram interval
VAR	variable name

Examples

Example 1: Univariate Analysis for Multiple Variables

Procedure features:
VAR statement

This example computes the univariate statistics for two variables.

Program

```
options nodate pageno=1 linesize=80 pagesize=72;
```

The data set STATEPOP contains information from the 1980 and 1990 U.S. Census on the population in metropolitan and nonmetropolitan areas. The 50 states and District of Columbia are divided into four geographic regions. The data are organized by state within each region. The metropolitan and nonmetropolitan population counts are stored in one observation for both census years. A DATA step"STATEPOP" on page 1534 creates the data set.

```
data statepop;
   input State $ citypop_80 citypop_90
         Noncitypop_80 Noncitypop_90 Region @@;
   label citypop_80='1980 metropolitan pop in millions'
         noncitypop_80='1980 nonmetropolitan pop in millions'
         citypop_90='1990 metropolitan pop in millions'
         noncitypop_90='1990 nonmetropolitan pop in million'
         region='Geographic region';
    datalines;
ME   .405   .443  .721  .785 1   NH  .535   .659  .386  .450  1
NY 16.144 16.515 1.414 1.475 1   NJ 7.365  7.730   .A    .A   1
PA 10.067 10.083 1.798 1.799 1   DE  .496   .553  .098  .113  2
       ...more lines of data...
IA  1.198  1.200 1.716 1.577 3   MO 3.314  3.491 1.603 1.626 3
MT   .189   .191  .598  .608 4   ID  .257   .296  .687  .711 4
HI   .763   .836  2.02  .272 4
;
```

The VAR statement specifies the analysis variables and their order in the output.

```
proc univariate data=statepop;
   var citypop_90 citypop_80;
```

The TITLE statement specifies a title.

```
   title 'United States Census of Population and Housing';
run;
```

Output

Univariate statistics for both analysis variables appear on separate pages. Because each population value is unique, the mode is missing.

By comparing the two sums in the Moments table, you find that the metropolitan population increased by 20 million (197.7 - 176.9) in ten years. By comparing the two medians in Basic Statistical Measures table or the Quantiles table, you find that the 1990 median metropolitan population increased to 2.423 million.

```
          United States Census of Population and Housing                1

                      The UNIVARIATE Procedure
    Variable:  CityPop_90  (1990 metropolitan pop in millions)

                              Moments

N                          51    Sum Weights                 51
Mean                 3.87701961    Sum Observations       197.728
Std Deviation        5.16465302    Variance            26.6736408
Skewness             2.87109259    Kurtosis             10.537867
Uncorrected SS       2100.27737    Corrected SS        1333.68204
Coeff Variation       133.21194    Std Error Mean      0.72319608


                   Basic Statistical Measures

        Location                      Variability

     Mean     3.877020     Std Deviation            5.16465
     Median   2.423000     Variance                26.67364
     Mode      .           Range                   28.66500
                           Interquartile Range      3.60000


                   Tests for Location: Mu0=0

      Test           -Statistic-    -----p Value------

      Student's t    t  5.360952    Pr > |t|    <.0001
      Sign           M      25.5    Pr >= |M|   <.0001
      Signed Rank    S       663    Pr >= |S|   <.0001


                    Quantiles (Definition 5)

                     Quantile      Estimate

                     100% Max        28.799
                     99%             28.799
                     95%             14.166
                     90%              9.574
                     75% Q3           4.376
                     50% Median       2.423
                     25% Q1           0.776
                     10%              0.257
                     5%               0.191
                     1%               0.134
                     0% Min           0.134


                    Extreme Observations

               -----Lowest----        -----Highest----

                Value      Obs          Value      Obs

                0.134       41         10.083        9
                0.152        3         12.023       18
                0.191       39         14.166       26
                0.221       36         16.515        7
                0.226       50         28.799       49
```

United States Census of Population and Housing 2

The UNIVARIATE Procedure
Variable: CityPop_80 (1980 metropolitan pop in millions)

Moments

N	51	Sum Weights	51
Mean	3.46847059	Sum Observations	176.892
Std Deviation	4.427991	Variance	19.6071043
Skewness	2.47255319	Kurtosis	7.3709192
Uncorrected SS	1593.89992	Corrected SS	980.355217
Coeff Variation	127.664078	Std Error Mean	0.62004276

Basic Statistical Measures

Location		Variability	
Mean	3.468471	Std Deviation	4.42799
Median	2.114000	Variance	19.60710
Mode	.	Range	22.77400
		Interquartile Range	3.21000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	5.593922	Pr > |t|	<.0001
Sign	M	25.5	Pr >= |M|	<.0001
Signed Rank	S	663	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	22.907
99%	22.907
95%	11.539
90%	9.039
75% Q3	3.885
50% Median	2.114
25% Q1	0.675
10%	0.234
5%	0.174
1%	0.133
0% Min	0.133

Extreme Observations

-----Lowest----		-----Highest----	
Value	Obs	Value	Obs
0.133	3	9.461	29
0.141	41	10.067	9
0.174	50	11.539	26
0.189	39	16.144	7
0.194	36	22.907	49

Example 2: Rounding an Analysis Variable and Identifying Extreme Values

Procedure features:

PROC UNIVARIATE statement options:
FREQ
NEXTROBS=
NEXTRVAL=
ROUND=
ID statement

Data set: STATEPOP on page 1418

This example
- rounds the values of an analysis variable
- generates a frequency table
- identifies extreme observations.

Rounding affects all statistical computations. For this example, when the round unit is 1, all the nonnegative values below .5 round to zero.

Program

```
options nodate pageno=1 linesize=80 pagesize=68;
```

FREQ produces a frequency table. ROUND=1 rounds each value to the nearest integer.

```
proc univariate data=statepop freq round=1 nextrobs=2
                nextrval=4;
```

The VAR statement specifies an analysis variable.

```
    var citypop_90;
```

The ID statement specifies the variables to identify extreme observations. The TITLE statement specifies a title.

```
    id region state;
    title 'United States Census of Population and Housing';
run;
```

Output

The output includes a message to indicate that the values are rounded to the nearest integer. The Extreme Observations table lists values of the ID variables, Region and State. Region 4 reports the lowest metropolitan populations, while region 1 and region 4 report the highest populations. The states with the four most extreme observations are AK, WY, NY, and CA.

United States Census of Population and Housing 1

The UNIVARIATE Procedure
Variable: CityPop_90 (1990 metropolitan pop in millions)
Values Rounded to the Nearest Multiple of 1

Moments

N	51	Sum Weights	51
Mean	3.8627451	Sum Observations	197
Std Deviation	5.23457585	Variance	27.4007843
Skewness	2.85345529	Kurtosis	10.3550751
Uncorrected SS	2131	Corrected SS	1370.03922
Coeff Variation	135.5144	Std Error Mean	0.73298723

Basic Statistical Measures

Location		Variability	
Mean	3.862745	Std Deviation	5.23458
Median	2.000000	Variance	27.40078
Mode	1.000000	Range	29.00000
		Interquartile Range	3.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	5.269867	Pr > |t|	<.0001
Sign	M	21.5	Pr >= |M|	<.0001
Signed Rank	S	473	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	29
99%	29
95%	14
90%	10
75% Q3	4
50% Median	2
25% Q1	1
10%	0
5%	0
1%	0
0% Min	0

Extreme Observations

Lowest				Highest			
Value	Region	State	Obs	Value	Region	State	Obs
0	4	AK	50	17	1	NY	7
0	4	WY	41	29	4	CA	49

The Extreme Values table lists the four lowest unique values and the four highest unique values. Because ties occur in the values, the frequency counts of the values are shown.
The Frequency Counts table lists the variable values, the frequencies, the percentages, and the cumulative percentages.

United States Census of Population and Housing 2

The UNIVARIATE Procedure
Variable: CityPop_90 (1990 metropolitan pop in millions)
Values Rounded to the Nearest Multiple of 1

Extreme Values

---------Lowest--------			-------Highest--------		
Order	Value	Freq	Order	Value	Freq
1	0	8	11	12	1
2	1	14	12	14	1
3	2	4	13	17	1
4	3	9	14	29	1

Frequency Counts

		Percents				Percents				Percents	
Value	Count	Cell	Cum	Value	Count	Cell	Cum	Value	Count	Cell	Cum
0	8	15.7	15.7	5	1	2.0	80.4	12	1	2.0	94.1
1	14	27.5	43.1	6	1	2.0	82.4	14	1	2.0	96.1
2	4	7.8	51.0	8	2	3.9	86.3	17	1	2.0	98.0
3	9	17.6	68.6	9	1	2.0	88.2	29	1	2.0	100.0
4	5	9.8	78.4	10	2	3.9	92.2				

Example 3: Computing Robust Estimators

Procedure features:

PROC UNIVARIATE statement options:
ROBUSTSCALE
TRIMMED=
WINSORIZED=

Data set: STATEPOP on page 1418

This example
- computes robust estimates of location
- computes two trimmed means
- computes a Winsorized mean.

Program

TRIMMED= computes two trimmed means after removing 6 observations and 25 percent of the observations. WINSORIZED= computes a Winsorized mean that replaces 10 percent of the observations.

```
options nodate pageno=1 linesize=80 pagesize=72;

proc univariate data=statepop robustscale trimmed=6 .25
                winsorized=.1;
```

The VAR statement specifies an analysis variable.

```
     var citypop_90;
```

The TITLE statement specifies a title.

```
   title 'United States 1990 Census of Population and Housing';
run;
```

Output

Because each value of population is unique, the mode is missing.

Both the trimmed and Winsorized means are smaller than the arithmetic mean. This may be due to the positive skewness of the data. PROC UNIVARIATE trims 6 observations or 11.76 percent of the data from the tails. When you request to trim 25 percent of the data, PROC UNIVARIATE trims 13 observations or 25.49 percent of the data from the tails. This is because the number of observations trimmed is the smallest integer greater than or equal to 12.75 (.25×51). Likewise, when you compute a Winsorized mean for 10 percent of the data (.1×51=5.1), PROC UNIVARIATE uses 6 observations or 11.76 percent of the data from the tails.

```
            United States 1990 Census of Population and Housing             1

                          The UNIVARIATE Procedure
          Variable:  CityPop_90  (1990 metropolitan pop in millions)

                                   Moments

       N                          51    Sum Weights                 51
       Mean                 3.87701961    Sum Observations       197.728
       Std Deviation        5.16465302    Variance            26.6736408
       Skewness             2.87109259    Kurtosis             10.537867
       Uncorrected SS       2100.27737    Corrected SS        1333.68204
       Coeff Variation       133.21194    Std Error Mean      0.72319608


                         Basic Statistical Measures

              Location                      Variability

         Mean     3.877020     Std Deviation            5.16465
         Median   2.423000     Variance                26.67364
         Mode      .           Range                   28.66500
                               Interquartile Range      3.60000


                         Tests for Location: Mu0=0

            Test           -Statistic-    -----p Value------

            Student's t    t  5.360952    Pr > |t|    <.0001
            Sign           M      25.5    Pr >= |M|   <.0001
            Signed Rank    S       663    Pr >= |S|   <.0001
```

Trimmed Means

Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits		DF
11.76	6	2.702231	0.535235	1.618705	3.785756	38
25.49	13	2.307000	0.438141	1.402721	3.211279	24

Trimmed Means

Percent Trimmed in Tail	t for H0: Mu0=0.00	Pr > |t|
11.76	5.048686	<.0001
25.49	5.265424	<.0001

Winsorized Means

Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits		DF
11.76	6	3.139588	0.536889	2.052713	4.226463	38

Winsorized Means

Percent Winsorized in Tail	t for H0: Mu0=0.00	Pr > |t|
11.76	5.847741	<.0001

United States 1990 Census of Population and Housing 2

The UNIVARIATE Procedure
Variable: CityPop_90 (1990 metropolitan pop in millions)

Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	3.600000	2.668683
Gini's Mean Difference	4.614921	4.089867
MAD	1.675000	2.483355
Sn	2.626105	2.673281
Qn	2.230788	2.171186

Quantiles (Definition 5)

Quantile	Estimate
100% Max	28.799
99%	28.799
95%	14.166
90%	9.574
75% Q3	4.376
50% Median	2.423
25% Q1	0.776
10%	0.257
5%	0.191
1%	0.134
0% Min	0.134

Extreme Observations

-----Lowest----		-----Highest----	
Value	Obs	Value	Obs
0.134	41	10.083	9
0.152	3	12.023	18
0.191	39	14.166	26
0.221	36	16.515	7
0.226	50	28.799	49

Example 4: Performing a Sign Test Using Paired Data

Procedure features:

PROC UNIVARIATE statement option:

ALPHA=
CIBASIC
CIPCTLDF
LOCCOUNT
MODES

Other features:

LABEL statement

This example

- computes difference scores for paired data
- lists all values of the mode
- examines the tests for location to determine if the median difference between scores is zero
- lists the number of observations less than, greater than, and equal to zero
- specifies the confidence levels for the confidence limits
- generates distribution-free confidence limits for the quantiles.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

The data set SCORE contains test scores for college students who took two tests and a final exam. ScoreChange contains the difference in the score between the first test and the second test.

```
data score;
   input Student $ Test1 Test2 Final @@;
   ScoreChange=test2-test1;
   datalines;
Capalleti  94 91 87  Dubose     51 65 91
Engles     95 97 97  Grant      63 75 80
Krupski    80 75 71  Lundsford  92 55 86
Mcbane     75 78 72  Mullen     89 82 93
Nguyen     79 76 80  Patel      71 77 83
Si         75 70 73  Tanaka     87 73 76
;
```

LOCCOUNT produces a Location Counts table. MODES produces a Modes table. ALPHA= specifies a 99 percent confidence limit as the default for all statistics. CIBASIC(ALPHA=.05) specifies a 95 percent confidence limit for the basic measures. CIPCTLDF produces distribution–free confidence limits for the quantiles.

```
proc univariate data=score loccount modes alpha=.01
                cibasic(alpha=.05) cipctldf;
```

The VAR statement specifies the analysis variable as the test scores differences.

```
   var scorechange;
```

The LABEL statement associates a label with the analysis variable for the duration of the PROC step. The TITLE statement specifies a title.

```
   label scorechange='Change in Test Scores';
   title 'Test Scores for a College Course';
run;
```

Output

PROC UNIVARIATE includes the variable label in the report. The report also provides a message to indicate that the lowest mode is shown in the Basic Statistical Measures table. The Modes table reports all the mode values.

The mean of -3.08 indicates an average decrease in test scores from Test1 to Test2. The 95 percent confidence limits (-11.56, 5.39), which includes 0, and the tests for location indicate that the decrease is not statistically significant.

The Tests for Location table includes three hypothesis tests. The Student's **t** statistic assumes that the data are approximately normally distributed. The sign test and signed rank test are nonparametric tests. The signed rank test requires a symmetric distribution. If the distribution is symmetric you expect a skewness value that is close to zero. Because the value -1.42 indicates some distribution skewness, examine the sign test to determine if the difference in test scores is zero. The large **p**-value (.7744) provides insufficient evidence of a difference in test score medians.

```
                    Test Scores for a College Course                      1

                        The UNIVARIATE Procedure
                Variable:  ScoreChange  (Change in Test Scores)

                                Moments

        N                          12    Sum Weights                12
        Mean                -3.0833333    Sum Observations          -37
        Std Deviation       13.3379727    Variance            177.901515
        Skewness            -1.4191368    Kurtosis            3.35291936
        Uncorrected SS            2071    Corrected SS        1956.91667
        Coeff Variation      -432.5829    Std Error Mean      3.85034106


                      Basic Statistical Measures

             Location                      Variability

          Mean     -3.08333     Std Deviation           13.33797
          Median   -3.00000     Variance               177.90152
          Mode     -5.00000     Range                   51.00000
                                Interquartile Range     10.50000

     NOTE: The mode displayed is the smallest of 2 modes with a count of 2.
```

Modes

Mode	Count
-5	2
-3	2

Basic Confidence Limits Assuming Normality

Parameter	Estimate	95% Confidence Limits	
Mean	-3.08333	-11.55788	5.39121
Std Deviation	13.33797	9.44856	22.64625
Variance	177.90152	89.27519	512.85267

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	-0.80079	Pr > |t|	0.4402
Sign	M	-1	Pr >= |M|	0.7744
Signed Rank	S	-8.5	Pr >= |S|	0.5278

Location Counts: Mu0=0.00

Count	Value
Num Obs > Mu0	5
Num Obs ^= Mu0	12
Num Obs < Mu0	7

Because PROC UNIVARIATE computes a symmetric confidence interval, some coverages for the confidence limits are less than 99 percent. In some cases, there are also insufficient data to compute a symmetric confidence interval and a missing value is shown. Use the TYPE=ASYMMETRIC option to increase the coverage and reduce the number of missing confidence limits.

```
                    Test Scores for a College Course                        2

                        The UNIVARIATE Procedure
                Variable:  ScoreChange  (Change in Test Scores)

                        Quantiles (Definition 5)

                         Quantile      Estimate

                         100% Max          14.0
                         99%               14.0
                         95%               14.0
                         90%               12.0
                         75% Q3             4.5
                         50% Median        -3.0
                         25% Q1            -6.0
                         10%              -14.0
                         5%               -37.0
                         1%               -37.0
                         0% Min           -37.0

                        Quantiles (Definition 5)

                 99% Confidence Limits    -------Order Statistics-------
Quantile           Distribution Free      LCL Rank  UCL Rank    Coverage

100% Max
99%                  6                14          10        12       11.34
95%                  6                14          10        12       44.01
90%                  2                14           8        12       71.32
75% Q3              -3                14           6        12       95.41
50% Median         -14                12           2        11       99.37
25% Q1             -37                -3           1         7       95.41
10%                -37                -5           1         5       71.32
5%                 -37                -7           1         3       44.01
1%                 -37                -7           1         3       11.34
0% Min

                            Extreme Observations

                       ----Lowest----        ----Highest---

                       Value      Obs        Value      Obs

                         -37        6            2        3
                         -14       12            3        7
                          -7        8            6       10
                          -5       11           12        4
                          -5        5           14        2
```

Example 5: Examining the Data Distribution and Saving Percentiles

Procedure features:

PROC UNIVARIATE statement options:

ALPHA=

CIBASIC
CIPCTLNORMAL
MU0=
NORMAL
PLOTS
PLOTSIZE=
OUTPUT statement

Other features:
PRINT procedure

Data set: SCORE on page 1428

This example

- specifies the confidence level for the confidence limits
- computes a lower confidence limit for the parameters
- computes two-sided confidence limits for the quantiles based on the assumption of normal data
- specifies the null hypothesis mean for the tests for locations
- tests the hypothesis that the data are normally distributed
- produces a stem-and-leaf plot, box plot, and normal probability plot and increase the plot size
- computes additional percentiles
- creates an output data set with percentiles
- prints the output data set.

Program

```
options nodate pageno=1 linesize=64 pagesize=58;
```

MU0= requests a test that the population mean equals 80. ALPHA= specifies a 90 percent confidence limit for all statistics. CIBASIC computes lower confidence limits for the basic measures. CIPCTLNORMAL computes two-sided confidence limits for the quantiles. NORMAL computes tests for normality. PLOTS requests plots of the data distribution. PLOTSIZE= specifies the number of rows to display the plot.

```
proc univariate data=score mu0=80 alpha=.1 cibasic(type=lower)
                cipctlnormal normal plots plotsize=26;
```

The VAR statement specifies the analysis variable.

```
   var final;
```

The OUTPUT statement creates the PCTSCORE data set with five variables. MEDIAN= saves the median. PCTLPTS= saves four percentiles. PCTLPRE= specifies a prefix name. PCTLNAME= specifies suffix names for the variables that contain the first three percentiles. The name of the variable that contains the 70th percentile uses the default suffix.

```
   output out=pctscore median=Median pctlpts=98 50 20 70
          pctlpre=Pctl_ pctlname=Top Mid Low;
   title 'Examining the Distribution of Final Exam Scores';
run;
```

PROC PRINT prints the PCTSCORE data set. The TITLE statement specifies a title.

```
proc print data=pctscore noobs;
   title1 'Quantile Statistics for Final Exam Scores';
   title2 'Output Data Set from PROC UNIVARIATE';
run;
```

Output

The estimate of the mean test score is 82.4, with a standard deviation of 8.6. The 90 percent lower confidence limit for the mean is 79.

The Tests for Location table includes three hypothesis tests. To determine whether the Student's **t** statistic is appropriate, you must determine if the data are approximately normally distributed. PROC UNIVARIATE calculates the Shapiro-Wilk W statistic because the sample size is below 2000. All **p**-values from the tests for normality are >0.15, which provides insufficient evidence to reject the assumption of normality. The probability plot also supports the assumption that the data are normal. Therefore, the **t** statistic appears appropriate. The **p**-value of .35 for this test provides insufficient evidence to reject the null hypothesis that the mean test score is 80.

Examination of the box plot, which is nonsymmetric, and the small sample size, which causes low power, make the sign test a more appropriate test of location. The **p**-value of .75 for this test provides insufficient evidence to reject the null hypothesis that the mean test score is 80.

```
        Examining the Distribution of Final Exam Scores        1

                    The UNIVARIATE Procedure
                        Variable:  Final

                             Moments

N                          12    Sum Weights                12
Mean                 82.4166667    Sum Observations          989
Std Deviation        8.59659905    Variance           73.9015152
Skewness             0.22597472    Kurtosis           -1.0846549
Uncorrected SS            82323    Corrected SS       812.916667
Coeff Variation     10.4306561    Std Error Mean     2.48162439


                    Basic Statistical Measures

          Location                     Variability

      Mean     82.41667     Std Deviation            8.59660
      Median   81.50000     Variance                73.90152
      Mode     80.00000     Range                   26.00000
                            Interquartile Range     14.50000


            Basic Confidence Limits Assuming Normality

            Parameter          Estimate    Lower 90% CL

            Mean               82.41667      79.03314
            Std Deviation       8.59660       6.85984
            Variance           73.90152      47.05738


                    Tests for Location: Mu0=80

         Test           -Statistic-    -----p Value------

         Student's t    t  0.973825    Pr > |t|    0.3511
         Sign           M         1    Pr >= |M|   0.7539
         Signed Rank    S         8    Pr >= |S|   0.4434


                      Tests for Normality

  Test                   --Statistic---    ------p Value------

  Shapiro-Wilk           W     0.952903    Pr < W       0.6797
  Kolmogorov-Smirnov     D     0.113328    Pr > D      >0.1500
  Cramer-von Mises       W-Sq  0.028104    Pr > W-Sq   >0.2500
  Anderson-Darling       A-Sq  0.212693    Pr > A-Sq   >0.2500
```

The three plots display the data distribution. The PLOTSIZE= option enlarges the plots so that you can easily see if the data are approximately normal.

```
Examining the Distribution of Final Exam Scores        2

            The UNIVARIATE Procedure
                Variable:  Final

           Quantiles (Definition 5)

                                90% Confidence Limits
Quantile      Estimate      Assuming Normality

100% Max          97.0
99%               97.0    96.30698     114.6289
95%               97.0    91.55028     105.9399
90%               93.0    88.89956     101.4163
75% Q3            89.0    84.12815      94.1623
50% Median        81.5    77.95996      86.8734
25% Q1            74.5    70.67102      80.7052
10%               72.0    63.41705      75.9338
5%                71.0    58.89343      73.2831
1%                71.0    50.20448      68.5264
0% Min            71.0


              Extreme Observations

      ----Lowest----        ----Highest---

      Value      Obs        Value      Obs

         71        5           86        6
         72        7           87        1
         73       11           91        2
         76       12           93        8
         80        9           97        3


Stem Leaf                     #             Boxplot
  96 0                        1                |
  94                                           |
  92 0                        1                |
  90 0                        1                |
  88                                          +-----+
  86 00                       2               |     |
  84                                          |     |
  82 0                        1               |  +  |
  80 00                       2               *-----*
  78                                          |     |
  76 0                        1               |     |
  74                                          +-----+
  72 00                       2                  |
  70 0                        1                  |
     ----+----+----+----+
```

```
        Examining the Distribution of Final Exam Scores        3

                    The UNIVARIATE Procedure
                        Variable:  Final

                         Normal Probability Plot
        97+                                                *++
           |                                             ++
           |                                          *++
           |                                        *+++
           |                                       ++
           |                                   * *+
           |                                  +++
           |                                +*
           |                             *++
           |                           +++
           |                         +++*
           |                      ++
           |                    +*+ *
        71+           * ++
              +----+----+----+----+----+----+----+----+----+----+----+
                   -2         -1          0         +1         +2
```

The PCTSCORE data set contains one observation. The median value in Median is equivalent to the 50th percentile in PCTL_MID.

```
              Quantile Statistics for Final Exam Scores           4
                 Output Data Set from PROC UNIVARIATE

      Median     Pctl_Top     Pctl_Mid     Pctl_Low     Pctl_70

       81.5         97          81.5          73           87
```

Example 6: Creating an Output Data Set with Multiple Analysis Variables

Procedure features:

PROC UNIVARIATE statement option:

NOPRINT

OUTPUT statement

VAR statement

Other features:

PRINT procedure

Data set: SCORE on page 1428

This example

- suppresses the reporting of univariate statistics
- computes additional percentiles for two variables

- creates an output data set with descriptive statistics and percentiles
- prints the output data set.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

NOPRINT suppresses all the tables of statistics.

```
proc univariate data=score noprint;
```

The VAR statement specifies the analysis variables and their order in the output.

```
   var test1 test2;
```

The OUTPUT statement creates the TESTSTAT data set with nine variables. MEAN= saves the mean for Test1 and Test2. STD= saves the standard deviation for Test1. PCTLPTS= calculates three percentiles and PCTLPRE= specifies prefix names for the analysis variables. PCTLNAME= specifies a suffix name for the 33.3 percentile.

```
   output out=teststat mean=MeanTest1 MeanTest2
          std=StdDeviationTest1
          pctlpts=33.3 66 99.9
          pctlpre=Test1_
          Test2_ pctlname=Low ;
run;
```

PROC PRINT prints the TESTSTAT data set. The TITLE statement specifies a title.

```
proc print data=teststat noobs;
   title1 'Univariate Statistics for Two College Tests';
   title2 'Output Data Set from PROC UNIVARIATE';
run;
```

Output

The TESTSTAT data set contains one observation with the mean for the two analysis variables and the standard deviation for the first analysis variable. The remaining six variables contain computed percentiles.

```
                  Univariate Statistics for Two College Tests                  1
                      Output Data Set from PROC UNIVARIATE

                     Std
 Mean     Mean    Deviation  Test1_             Test1_  Test2_             Test2_
Test1    Test2      Test1      Low   Test1_66    99_9     Low   Test2_66   99_9

79.25  76.1667    13.3152     75        87        95      73        77        97
```

Example 7: Creating Schematic Plots and an Output Data Set with BY Groups

Procedure features:

PROC UNIVARIATE statement options:

NEXTROBS=
PLOT
PLOTSIZE=

BY statement

OUTPUT statement

Other features:

FORMAT statement

FORMAT procedure

PRINT procedure

SORT procedure

Data set: STATEPOP on page 1418

This example

- creates a data set with observations that are separated by census year
- sorts the data set by geographic region and census year
- calculates univariate statistics and produces a stem-and-leaf plot, box plot, and normal probability plot for each BY group
- creates schematic plots to compare the BY groups
- creates an output data set with descriptive statistics and percentiles
- prints the output data set.

Program

```
options nodate pageno=1 linesize=120 pagesize=80;
```

PROC FORMAT creates a format to identify regions with a character value.

```
proc format;
   value Regnfmt 1='Northeast'
D                 2='South'
                  3='Midwest'
                  4='West';
run;
```

The METROPOP data set contains one variable, Populationcount, with the metropolitan and nonmetropolitan population counts. DECADE indicates the census year for the observation. The OUTPUT statements create two observations for each state and decade combination.

```
data metropop;
   set statepop;
   keep Region Decade Populationcount;
   label PopulationCount='US Census Population (millions)'
         Decade='Census year';
   decade=1980;
   populationcount=sum(citypop_80,noncitypop_80);
   output;
   decade=1990;
   populationcount=sum(citypop_90,noncitypop_90);
   output;
```

PROC SORT sorts the observations by Region and Decade.

```
proc sort data=metropop;
   by region decade;
run;
```

NEXTROBS= suppresses the Extreme Observations table. PLOTS produces plots that show the data distribution. PLOTSIZE= specifies the number of rows used to display the plots.

```
proc univariate data=metropop nextrobs=0
                plots plotsize=20 ;
```

The VAR statement specifies the analysis variable.

```
   var populationcount;
```

The BY statement produces a separate section of the report for each BY group and prints a heading above each one.

```
   by region decade;
```

The OUTPUT statement creates the CENSTAT data set with six variables and eight observations. SUM= saves the sum. MEAN= saves the mean. STD= saves the standard deviation. PCTLPTS= calculates three percentiles. PCTLPRE= specifies the prefix name.

```
   output out=censtat sum=PopulationTotal mean=PopulationMean
          std=PopulationStdDeviation pctlpts=50 to 100 by 25
          pctlpre=Pop_ ;
```

The FORMAT statement assigns a format to Region. The output data set contains the formatted values of Region. The TITLE statement specifies a title.

```
   format region regnfmt.;
   title 'United States Census of Population and Housing';
run;
```

PROC PRINT prints the CENSTAT data set.

```
proc print data=censtat;
   title1 'Statistics for Census Data by Decade and Region';
   title2 'Output Dataset From PROC UNIVARIATE';
run;
```

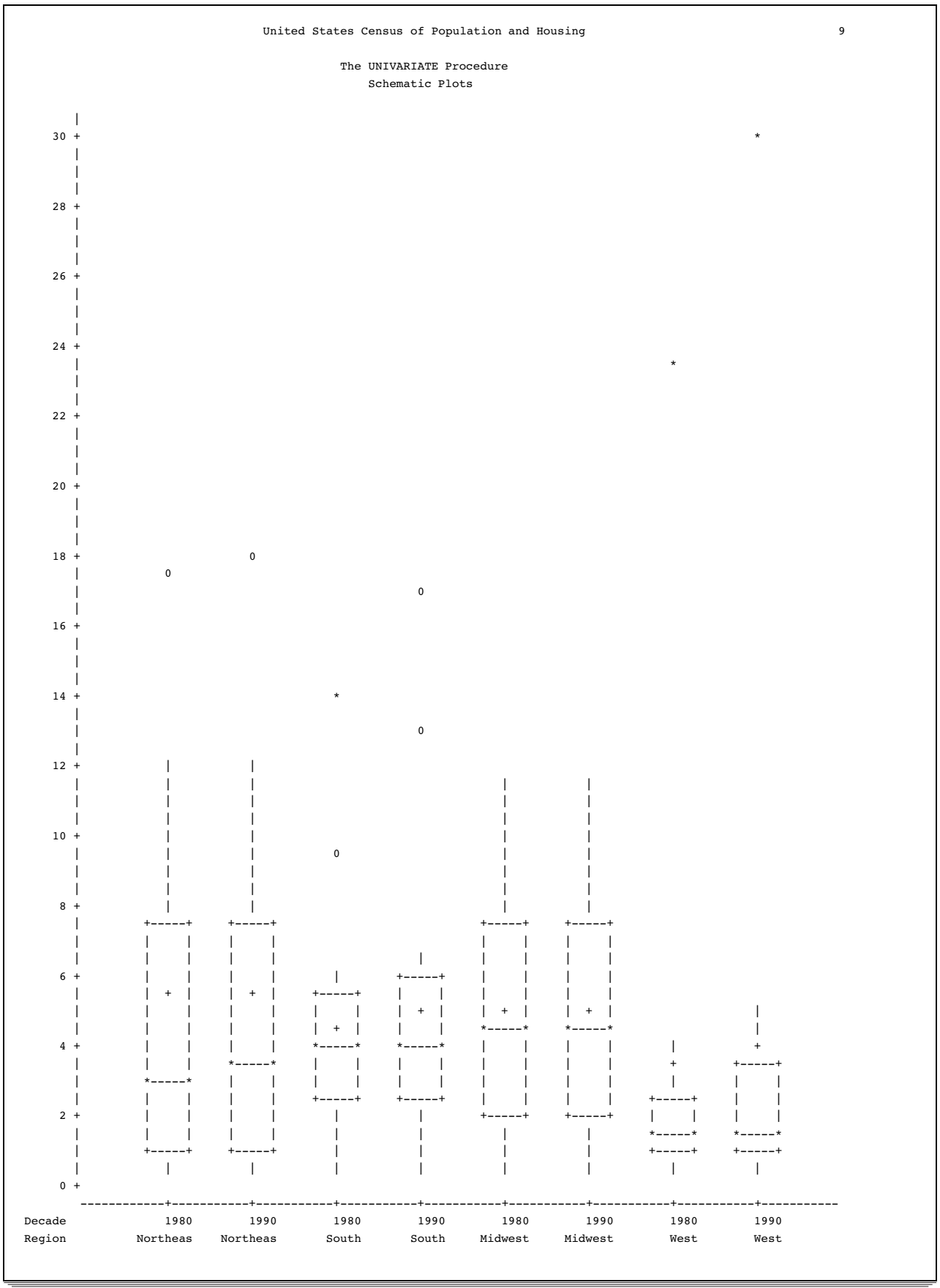
Output

The UNIVARIATE procedure output that is shown does not include univariate statistics for each BY group. Only univariate statistics for the first BY group and schematic plots for all BY groups are shown.

The BY statement requests separate reports for each BY group. The first report contains univariate statistics for the 1980 Census, Northeast region.

```
                                United States Census of Population and Housing                                1

------------------------------------ Geographic region=Northeast Census year=1980 -------------------------------------

                                      The UNIVARIATE Procedure
                              Variable:  PopulationCount  (US Census Population (millions))

                                           Moments

                          N                          9    Sum Weights                  9
                          Mean               5.45977778    Sum Observations        49.138
                          Std Deviation      5.92525457    Variance            35.1086417
                          Skewness           1.26661463    Kurtosis            0.84864338
                          Uncorrected SS     549.151694    Corrected SS        280.869134
                          Coeff Variation    108.525563    Std Error Mean      1.97508486


                                      Basic Statistical Measures

                                Location                    Variability

                             Mean     5.459778     Std Deviation            5.92525
                             Median   3.108000     Variance                35.10864
                             Mode      .           Range                   17.04700
                                                   Interquartile Range      6.41800


                                      Tests for Location: Mu0=0

                           Test           -Statistic-    -----p Value------

                           Student's t    t  2.764326    Pr > |t|    0.0245
                           Sign           M       4.5    Pr >= |M|   0.0039
                           Signed Rank    S      22.5    Pr >= |S|   0.0039


                                       Quantiles (Definition 5)

                                        Quantile      Estimate

                                        100% Max        17.558
                                        99%             17.558
                                        95%             17.558
                                        90%             17.558
                                        75% Q3           7.365
                                        50% Median       3.108
                                        25% Q1           0.947
                                        10%              0.511
                                        5%               0.511
                                        1%               0.511
                                        0% Min           0.511


Stem Leaf                     #          Boxplot                              Normal Probability Plot
  16 6                        1             0                  17+                                      *  ++++
  14                                                           |                                      +++
  12                                                           |                                  +++
  10 9                        1             |                  |                              +*++
   8                                        |                  9+                          +++
   6 4                        1          +-----+               |                         ++++ *
   4 7                        1          |  +  |               |                      +++  *
   2 1                        1          *-----*               |                   +++  *
   0 5991                     4          +-----+               1+          *     *+++*  *
     ----+----+----+----+                                        +----+----+----+----+----+----+----+----+----+----+
                                                                    -2        -1         0        +1        +2
```

The BY statement and a PLOT option in PROC statement produce schematic plots on the last page of the report. Use the graph to compare the data distribution for each region-year combination.



The CENSTAT data set includes the BY variables Region and Decade and contains eight observation, one for each BY group.

Statistics for Census Data by Decade and Region 10
Output Dataset From PROC UNIVARIATE

Obs	Region	Decade	Population Mean	Population Std Deviation	Population Total	Pop_50	Pop_75	Pop_100
1	Northeast	1980	5.45978	5.92525	49.138	3.1080	7.365	17.558
2	Northeast	1990	5.64556	6.00833	50.810	3.2880	7.730	17.990
3	South	1980	4.43329	3.32034	75.366	3.8940	5.347	14.225
4	South	1990	5.02647	4.20752	85.450	4.0410	6.187	16.987
5	Midwest	1980	4.90567	3.75037	58.868	4.3910	7.376	11.428
6	Midwest	1990	4.97242	3.75702	59.669	4.6335	7.420	11.431
7	West	1980	3.32154	6.21703	43.180	1.3030	2.717	23.667
8	West	1990	4.06000	7.83953	52.780	1.5150	3.294	29.760

Example 8: Fitting Density Curves

Procedure features:

PROC UNIVARIATE statement options:

NOPRINT

HISTOGRAM statement options:

CBARLINE=
CFILL=
EXP
FILL
L=
MIDPOINTS=
NOPRINT
NORMAL

VAR statement

Other features:

GOPTIONS statement
RANNOR function
RANEXP function

This example

- creates a sample of 100 observations from a normal distribution and an exponential distribution
- suppresses the tables of descriptive statistics
- creates histograms with superimposed density curves for the normal and exponential distributions
- requests goodness-of-fit tests for a fitted exponential distribution
- specifies the midpoints for histogram intervals
- requests graphical enhancements that change plot colors and line types.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

The GOPTIONS statement sets the graphics environment to control the appearance of graphics elements. HTITLE= and HTEXT= specify text height. FTEXT= and FTITLE= specify the font.*

```
goptions htitle=4 htext=3 ftext=swissb ftitle=swissb;
```

The data set DISTRDATA contains two variables and 100 observations. The RANNOR function creates a random variate from a normal distribution with a mean of 50 and standard deviation of 10 that is stored in the Normal_x variable. The RANEXP function creates a random variate from a exponential distribution that is stored in the Exponential_x variable.

```
data distrdata;
   drop n;
   label Normal_x='Normal Random Variable'
         Exponential_x='Exponential Random Variable';
   do n=1 to 100;
      Normal_x=10*rannor(53124)+50;
      Exponential_x=ranexp(18746363);
      output;
   end;
run;
```

NOPRINT suppresses the tables of statistics that the PROC UNIVARIATE statement creates. The VAR statement specifies the analysis variable.

```
proc univariate data=distrdata noprint;
   var Normal_x;
```

The HISTOGRAM statement creates a histogram for the analysis variable Normal_x. The NORMAL option superimposes the fitted density curve for a normal distribution. NOPRINT suppresses the tables of statistics that summarize the fitted density curve. The CBARLINE= option specifies the color to outline the histogram bars.

```
histogram Normal_x /normal(noprint) cbarline=grey ;
```

The TITLE statement specifies a title.

```
title '100 Obs Sampled from a Normal Distribution';
run;
```

Another PROC step will execute so that output displays a new customized title. The VAR statement specifies the analysis variable.

```
proc univariate data=distrdata noprint;
  var Exponential_x;
```

The HISTOGRAM statement creates a histogram for the analysis variable Exponential_x. The EXP option superimposes a fitted density curve for an exponential distribution . The FILL option specifies to fill the area under the exponential density curve with the CFILL= color. The L= option specifies a distinct line type for the density curve. The MIDPOINTS= option specifies a list of values to use as bin midpoints.

* For additional information about the GOPTIONS statement, see *SAS/GRAPH Software: Reference*.

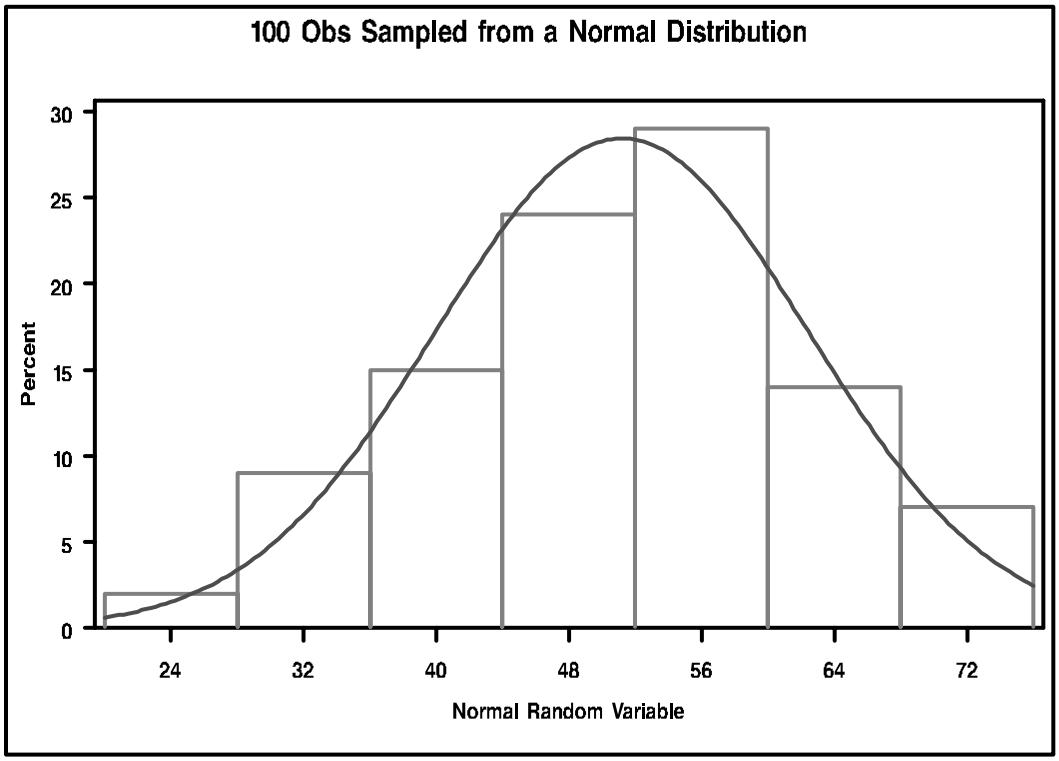
```
histogram /exp(fill l=3) cfill=yellow midpoints=.05 to 5.55 by .25;
```

The TITLE statement specifies a title.

```
title '100 Obs Sampled from an Exponential Distribution';
run;
```

Output

Figure 41.4 A Histogram Superimposed with Normal Curve

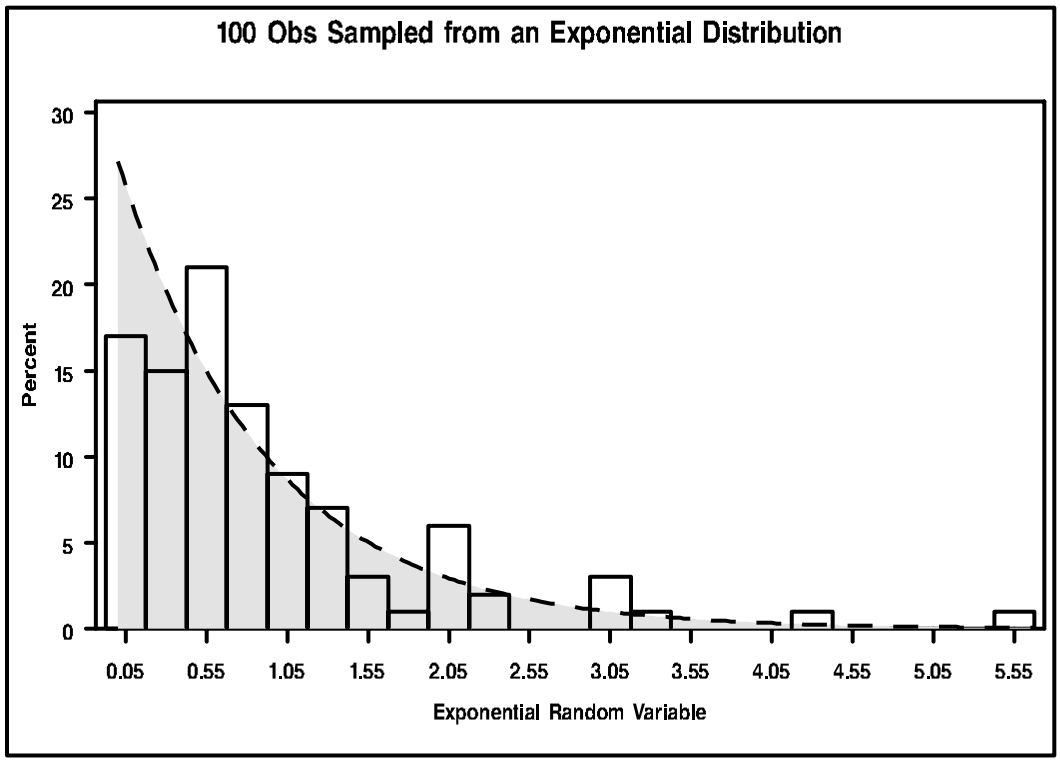


The output includes parameters estimates for the exponential curve. The exponential parameter threshold parameter θ is 0 because the THETA= option was omitted. A maximum likelihood estimate is computed for the scale parameter σ.

PROC UNIVARIATE provides three goodness-of-fit tests for the exponential distribution that are based on the empirical distribution function. The **p**–values for the exponential distribution are larger than the usual cutoff values of 0.05 and 0.10, which indicates not to reject the null hypothesis that the data are exponentially distributed.

```
100 Obs Sampled from an Exponential Distribution              1

          The UNIVARIATE Procedure
     Fitted Distribution for Exponential_x

     Parameters for Exponential Distribution

          Parameter   Symbol   Estimate

          Threshold   Theta           0
          Scale       Sigma    0.919698
          Mean                 0.919698
          Std Dev              0.919698


     Goodness-of-Fit Tests for Exponential Distribution

Test                  ---Statistic----   -----p Value-----

Kolmogorov-Smirnov    D     0.05860511   Pr > D     >0.150
Cramer-von Mises      W-Sq  0.05537161   Pr > W-Sq  >0.250
Anderson-Darling      A-Sq  0.33426909   Pr > A-Sq  >0.250
```

Figure 41.5 A Histogram Superimposed with an Exponential Curve



Example 9: Displaying a Reference Line on a Normal Probability Plot

Procedure features:

PROC UNIVARIATE statement options:

NOPRINT

INSET statement options:

FORMAT=
HEADER=
POSITION=
REFPOINT=
statistical-keyword

PROBPLOT statement options:

MU=
NORMAL
PCTLMINOR
SIGMA=

VAR statement

Other features:

GOPTIONS statement

SYMBOL statement

Data Set: DISTRDATA on page 1445

This example

- suppresses the tables of descriptive statistics
- creates a normal probability plot
- requests a diagonal reference line that corresponds to the normal distribution with estimated parameters μ and σ
- displays minor tick marks between major tick marks on the percentile axis
- enhances the plot by insetting a table of summary statistics
- requests graphical enhancements that change symbol type and text font.

Program

The GOPTIONS statement sets the graphics environment to control the appearance of graphics elements. HTITLE= and HTEXT= specify text height. FTEXT= and FTITLE= specify the font.*

```
goptions htitle=4 htext=3 ftext=swissb ftitle=swissb;
```

The SYMBOL statement defines the characteristics of the symbol that appears in the plot. VALUE= specifies a star for the plot symbol. By default, the plot symbol is the plus sign (+).

```
symbol value=star;
```

NOPRINT suppresses the tables of statistics that the PROC UNIVARIATE statement creates. The VAR statement specifies the analysis variable.

* For additional information about the GOPTIONS statement, see *SAS/GRAPH Software: Reference*.

```
proc univariate data=distrdata noprint;
  var Normal_x;
```

The PROBPLOT statement creates a normal probability plot for the analysis variable Normal_x. The NORMAL option superimposes a reference line that corresponds to the normal distribution by using estimated parameters for MU= and SIGMA=. PCTMINOR specifies that minor tick marks that appear between the major tick marks on the horizontal axis.

```
probplot normal_x /normal(mu=est sigma=est) pctlminor;
```

The INSET statement insets a table on the plot. The keywords MEAN and STD request that the mean and standard deviation display. FORMAT= specifies to use a format of field width 3. HEADER= displays a header at the top of the inset. POSITION= specifies to use axis percentage coordinates to position the inset. REFPOINT= specifies to place the bottom right corner of the inset 95% of the way across the horizontal axis and 5% of the way up the vertical axis.

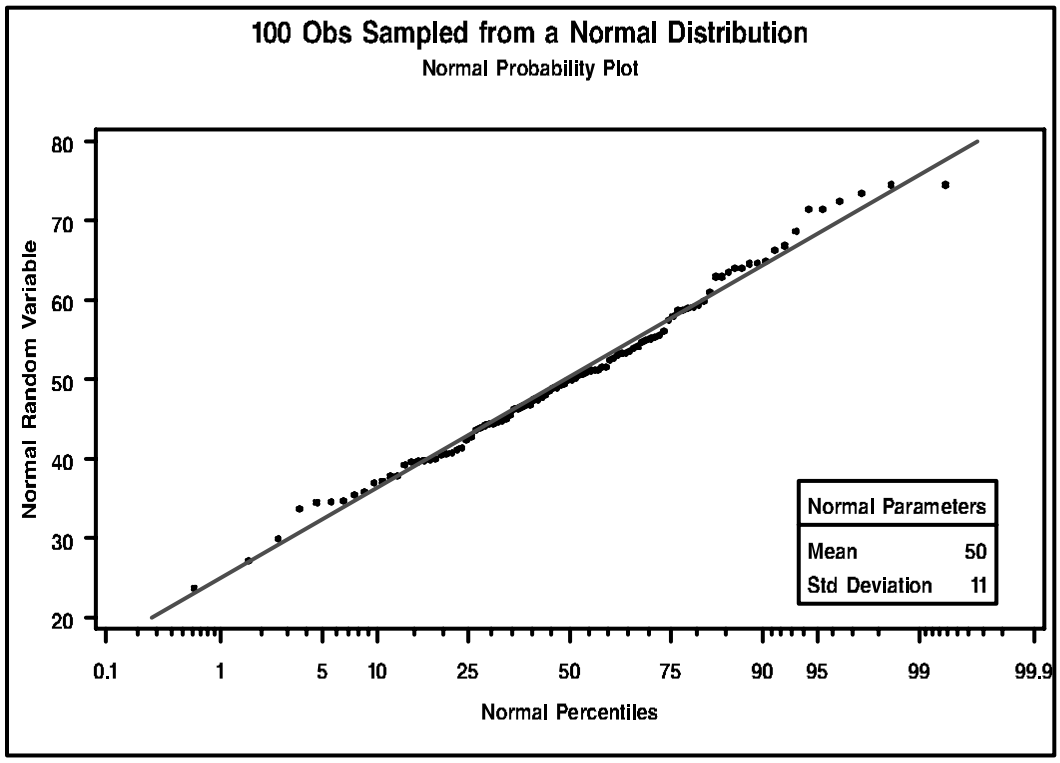
```
  inset mean std / format=3.0 header='Normal Parameters'
                   position=(95,5) refpoint=br;
```

The TITLE statements specify a title.

```
title1 '100 Obs Sampled from a Normal Distribution';
title2 'Normal Probability Plot';
run;
```

Output

Figure 41.6 Normal Probability Plot with a Normal Reference Line and a Customized Inset



Example 10: Creating a Two-Way Comparative Histogram

Procedure features:

PROC UNIVARIATE statement options:

NOPRINT

CLASS statement options:

ORDER=

HISTOGRAM statement options:

CFILL=
INTERTILE=
MIDPOINTS=
NCOLS=
NROWS=
VAXIS=
VAXISLABEL=
VSCALE=

INSET statement options:

FONT=
HEIGHT=
NOFRAME
POSITION=
statistical-keyword

VAR statement

Other features:

FORMAT statement

FORMAT procedure

GOPTIONS statement

Data set: METROPOP on page 1440

- □ suppresses the tables of descriptive statistics
- □ specifies two classification variables
- □ specifies the order of the component histograms
- □ creates a two-way comparative histogram with a specified number of rows and columns
- □
- □ specifies the distance between the component histogram tiles
- □ specifies the scale, values, and labels of the vertical axis
- □ specifies the midpoints for histogram intervals
- □ enhances the component histograms by insetting a table of summary statistics
- □ requests graphical enhancements that change fill color and font types.

Program

The GOPTIONS statement sets the graphics environment to control the appearance of graphic elements. HTITLE= and HTEXT= specify text height. FTEXT= and FTITLE= specify the font.*

```
goptions htitle=4 htext=3 ftext=swiss ftitle=swiss;
```

PROC FORMAT creates a format to identify regions with a character value.

```
proc format;
   value Regnfmt 1='Northeast'
                 2='South'
                 3='Midwest'
                 4='West';
run;
```

NOPRINT suppresses the tables of statistics that the PROC UNIVARIATE statement creates. The VAR statement specifies the analysis variable.

```
proc univariate data=metropop noprint;
    var populationcount;
```

The CLASS statement specifies Region and Decade as the classification variables. PROC UNIVARIATE produces a component histogram for each level (distinct combination of values) of these variables. ORDER= orders the classification levels by the frequency of Decade so that the year with greatest population count displays first.

```
   class region decade(order=freq);
```

The HISTOGRAM statement creates a two–way comparative histogram for the analysis variable PopulationCount. NROWS= and NCOLS= specify a 4 × 2 arrangement for the tiles. INTERTILE= inserts a space of one percentage screen unit between the tiles. CFILL= specifies a fill color for the histogram bars. VSCALE= requests the vertical axis scale in units of the number of observations per data unit. VAXIS= specifies the tick mark labels and VAXISLABEL= specifies a label for the vertical axis. MIDPOINTS= specifies a list of values to use as bin midpoints. FONT= requests a software font for the text.

```
    histogram /nrows=4 ncols=2 intertile=1 cfill=cyan vscale=count
              vaxis=0 4 8 12 vaxislabel='No. of States'
              midpoints=0 to 30 by 5;
```

* For additional information about the GOPTIONS statement, see *SAS/GRAPH Software: Reference*.

The INSET statement insets a table directly on each component histogram with the sum of PopulationCount. SUM= requests a customized label and a field width of five and two decimal places for the sum statistic. NOFRAME suppresses the frame around the inset table. POSITION= specifies a compass point to position the inset. HEIGHT= specifies the height of the text. FONT= requests a software font for the text.

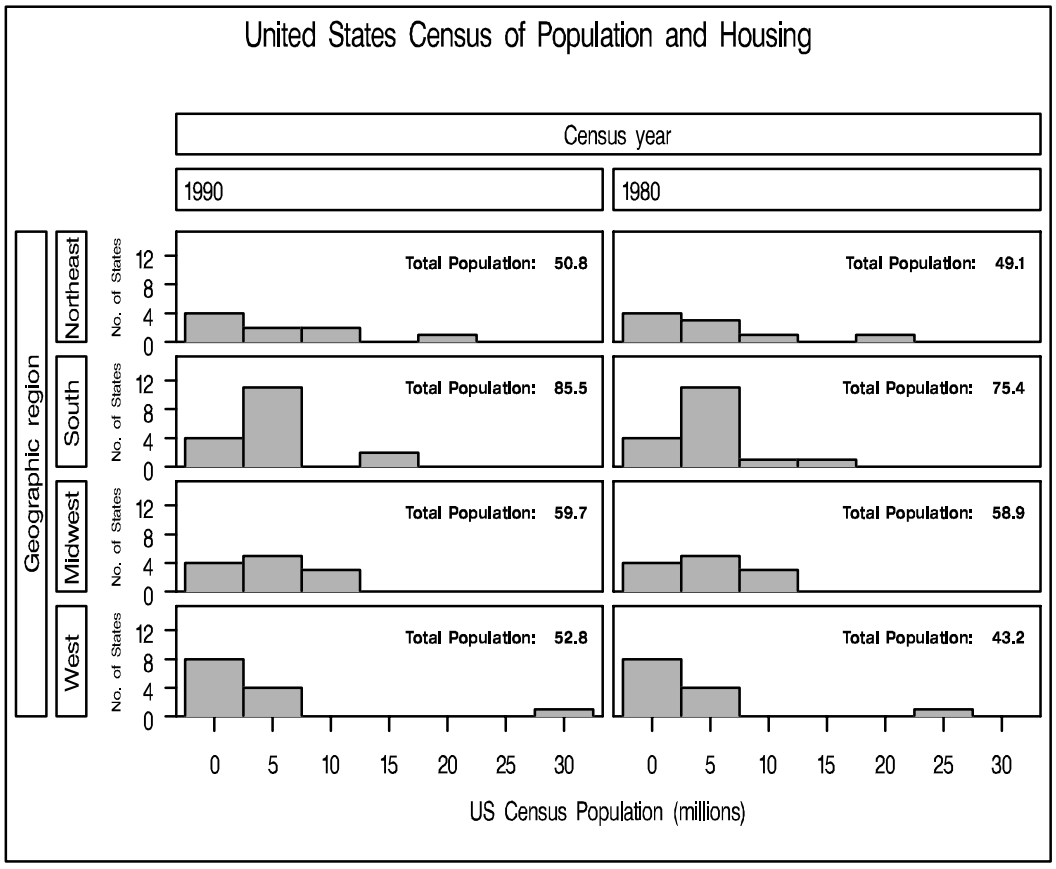
```
inset sum='Total Population:' (4.1) / noframe position=ne
                                      height=2 font=swissxb;
```

The FORMAT statement assigns a format to Region. The TITLE statement specifies a title.

```
   format region regnfmt.;
   title 'United States Census of Population and Housing';
run;
```

Output

Figure 41.7 Two-way Comparative Histogram



References

Blom, G. (1958), *Statistical Estimates and Transformed Beta Variables*, New York: John Wiley & Sons, Inc.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.

Conover, W.J. (1980), *Practical Nonparametric Statistics*, 2nd Edition, New York: John Wiley & Sons, Inc.

Croux, C. and Rousseeuw, P.J. (1992), "Time-Efficient Algorithms for Two Highly Robust Estimators of Scale," *Computational Statistics*, Volume 1, 411-428.

D'Agostino, R.B. and Stephens, M.A. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker, Inc.

Dixon, W.J. and Tukey, J.W. (1968), "Approximate Behavior of the Distribution of Winsorized *t* (Trimming/Winsorization 2)," *Technometrics*, 10, 83-98.

Frigge, M., Hoaglin, D.C., and Iglewicz, B. (1989), "Some Implementations of the Boxplot," *The American Statistician*, 43:1, 50–54.

Friendly, M. (1991) *SAS System for Statistical Graphics, First Edition*, Cary, NC: SAS Institute Inc.

Hahn, G.J. and Meeker, W. Q. (1991) *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley & Sons, Inc.

Hampel, F.R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383-393.

Iman, R.L. (1974), "Use of a t-statistic as an Approximation to the Exact Distribution of the Wilcoxon Signed Ranks Test Statistic," *Communications in Statistics*, 3, 795–806.

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions, Volume 1*, New York: John Wiley & Sons, Inc.

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1995), *Continuous Univariate Distributions, Volume 2*, New York: John Wiley & Sons, Inc.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day, Inc.

Odeh, R.E. and Owen, D.B. (1980), *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, New York: Marcel Dekker, Inc.

Owen, D.B. and Hua, T.A. (1977), "Tables of Confidence Limits on the Tail Area of the Normal Distribution," *Communication and Statistics, Part B – – Simulation and Computation*, 6, 285–311.

Rousseeuw, P.J. and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*. 88, 1273-1283.

Royston, J.P. (1992), "Approximating the Shapiro-Wilk's W-Test for Non-normality," *Statistics and Computing*, 2, 117–119.

Royston, J.P. (1982), "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples," *Applied Statistics*, 31, 115–124.

Shapiro, S.S. and Wilk, M.B. (1965), "An Analysis of Variance Test for Normality (complete samples)," *Biometrika*, 52, 591–611.

Schlotzhauer, S.D. and Littell, R.C. (1997) *SAS System for Elementary Statistical Analysis, Second Edition*, Cary, NC: SAS Institute Inc.

Stephens, M.A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730–737.

Terrell, G.R. and Scott, D.W. (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, 80, 209–214.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, Massachusetts: Addison-Wesley.

Tukey, J.W. and McLaughlin, D.H. (1963), "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/ Winsorization 1," *Sankhya A*, 25, 331-352.

U.S. Bureau of the Census (1994), *Statistical Abstract of the United States: 1994 (114th Edition)*, Washington, D.C.: U.S. Government Printing Office.

.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS® Procedures Guide, Version 8,* Cary, NC: SAS Institute Inc., 1999. 1729 pp.

SAS® Procedures Guide, Version 8

ISBN 1–58025–482–9

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999