

Data Summarization Methods in Base SAS Procedures

Lynne Bresler, SAS Institute, Inc, Cary, NC

ABSTRACT

Various Base SAS procedures compute an array of statistics to summarize your data. The types of statistical reports these procedures generate differ. With Version 8 enhancements, PROC UNIVARIATE now produces high-resolution displays of histograms, comparative histograms, probability plots, and quantile-quantile plots along with an optional table of descriptive statistics. The histograms can also include overlays of fitted density curves or kernel density estimates to examine the underlying data distribution. This paper compares how PROC MEANS, PROC TABULATE, and PROC UNIVARIATE generate descriptive statistics and describes how to use graphical displays to summarize your data.

INTRODUCTION

Base SAS® software provides several reporting procedures that generate summary statistics. This paper examines three commonly used procedures that summarize data. You are shown the different types of reports that PROC MEANS, PROC TABULATE, and PROC UNIVARIATE generate. Finally, you learn about the graphics statements that are available in Version 8 of Base SAS to examine the underlying data distribution.

To use the graphics statements in PROC UNIVARIATE to generate high-resolution histograms, probability plots, and quantile-quantile plots, you must license SAS/Graph® Software. You then have the ability to enhance these graphical displays by inseting a table with summary statistics that the procedure computes.

SELECTING A PROCEDURE

The exact procedure you select to summarize your data will depend on which statistics you want to show and how you want to organize the report. You also need to consider how quickly you expect to complete the analysis. PROC MEANS and PROC UNIVARIATE are simple to use and automatically produce a report of summary statistics. However, unless you use the Output Delivery System (ODS), you have limited control over the report layout. PROC TABULATE requires some understanding of complex syntax before you can construct a table with the statistics of interest. However, PROC TABULATE is a more appropriate choice when you want to design sophisticated tabular reports that summarize the data.

Table 1 in the Appendix displays a list of the statistics that the base SAS procedures commonly compute. You will find this table in Appendix 1, *SAS Elementary Statistics Procedures*, of the *SAS Procedures Guide* along with statistical notation, formulas, and standard keywords for these statistics. Documentation for the individual procedures discusses statistical concepts you may need to interpret the procedure output.

The MEANS procedure easily computes summary statistics for numeric variables across all observations. The report provides a concise summary of the data. When you specify classification variables with the CLASS statement, PROC MEANS computes statistics within groups of observations. The default report contains the number of observations, the mean, the standard deviation, the minimum value, and the maximum value. By specifying statistical keywords in the PROC MEANS statement, you control which statistics are in the report. However, you have a limited ability to format the report.

The UNIVARIATE procedure provides the most comprehensive summary of numeric data. By default, PROC UNIVARIATE produces a report of descriptive statistics based on moments, which includes skewness, kurtosis, and the coefficient of variation. The report also shows the mode, quantiles or percentiles (such as the median), tests for location, and information on the extreme values. Additional options in the PROC UNIVARIATE statement allow you to request robust estimates of location and scale, confidence limits, frequency tables, data distribution plots, and tests for normality. Unless you use ODS to customize a template, you have very little control over the format for statistics.

PROC TABULATE controls the layout of the statistics in a table while providing many of the same statistics that PROC MEANS, FREQ, REPORT, and UNIVARIATE compute. The tables you create range from simple to highly customized. The variables and statistics that define the pages, rows, and columns of the table determine how each table cell is calculated. After classifying the variables, you can create various hierarchical or nested groupings of the variables. Then the statistic associated with each cell is calculated on values from all observations in a category. The beauty of PROC TABULATE is the diversity of options that allow you to customize the labels and formats for the variables, statistics, and table headers.

CALCULATING QUANTILES

For large sample sizes the time required to calculate quantiles, which includes the median, is proportional to $n \log(n)$. Thus, a procedure like UNIVARIATE that automatically calculates quantiles can require more time than the other data summarization procedures. Furthermore, because data are held in memory, the procedure uses more storage space to perform its computations. By default, PROC MEANS and PROC TABULATE require less memory because they do not automatically compute quantiles. These procedures also provide the ability to use a fixed-memory quantiles estimation method that is usually less memory intense.

When you request a quantile statistic in PROC MEANS or PROC TABULATE, you can use the QMETHOD= option to specify the method that the procedure uses to compute quantiles. Two methods are available.

- OS which uses ordered statistics where all the data are read into memory and sorted by the unique values.
- P2 which uses a fixed space where all the data are accumulated into a fixed sample size to approximate the quantile.

QMETHOD=OS is the default and is the same method that PROC UNIVARIATE uses to compute quantiles. To compute weighted quantiles, you must use PROC UNIVARIATE or QMETHOD=OS.

QMETHOD=P2 is based on the piecewise-parabolic (P^2) algorithm developed by Jain and Chlamtac (1985). P^2 is a one-pass algorithm that is more efficient for large data sets because it requires a fixed amount of memory for each level within the type of each variable. However, based on simulation studies, estimates of some quantiles (P1, P5, P95, P99) may not be reliable for all data sets especially those with heavily tailed or skewed distributions.

CREATING GRAPHICAL DISPLAYS

PROC UNIVARIATE now supports high-resolution graphical displays. You can generate histograms and comparative histograms and optionally superimpose fitted probability density curves for various distributions and kernel density estimates. You can generate quantile-quantile plots, probability plots, and the corresponding comparative plots to compare a data distribution with a theoretical distribution. You also have the ability to inset summary statistics in the graphical displays.

The new graphics statements are

- HISTOGRAM statement
- PROBPLOT statement
- QQPLOT statement
- CLASS statement
- INSET statement

The HISTOGRAM statement creates a high-resolution graphics display of a histogram and optionally includes parametric and nonparametric density curve estimates. You can display density curves for several fitted theoretical distributions (beta, exponential, gamma, lognormal, normal, and Weibull), request goodness-of-fit tests for fitted distributions, and display kernel density estimates on histograms. You can specify the parameters of a distribution or request that PROC UNIVARIATE determine the parameter values. For most distributions, the default parameter values are maximum likelihood estimates.

The PROBPLOT statement creates a high-resolution graphics display of a probability plot. You can use the probability plot to compare the ordered variable values with the percentiles of a specified theoretical distribution.

The QQPLOT statement creates a graphical display of a quantile-quantile plot (Q-Q plot). You can use the Q-Q plot to compare the ordered variable values with quantiles of a specified theoretical distribution. The theoretical distributions you can select are beta, exponential, gamma, lognormal, normal, two-parameter Weibull, or three-parameter Weibull.

You can use the CLASS statement with a HISTOGRAM, PROBPLOT, or QQPLOT statement to create one-way and two-way high-resolution comparative plots. When you use a single class variable, PROC UNIVARIATE displays an array of component plots (stacked or side-by-side), for each level of the class variable. When you use two class variables, PROC UNIVARIATE displays a matrix of component plots, one for each combination of the levels of the class variables.

The INSET statement places a box or table of summary statistics, called an *inset*, directly in the graphical display. The inset can display the statistics that PROC UNIVARIATE calculates or display values that you provide in a SAS data set. The INSET statement does not produce the graphical display. You must first specify a HISTOGRAM, PROBPLOT, or QQPLOT statement. You can use options in the INSET statement to specify the position of the inset, a header for the inset, and various graphical enhancements, such as background colors, text colors, text height and text.

PRODUCING A SUMMARY REPORT

The PROC steps that follow summarize the data set CLINIC_STUDY. The data set has 360 randomly generated values for a hypothetical clinical trial. The variables are

Patient

a character string that uniquely identifies the patient.

Gender

a number that identifies the gender of the patient. A male is 1 and a female is 2.

Treatment

a number that classifies the type of treatment the patient received (placebo, drug, diet, or drug and diet).

Height

the height of the patient in inches. The height for males is 64 to 76 and for females is 58 to 70.

Weight

the weight of the patient in pounds. The weight for males is 120 to 280 and for females is 105 to 210.

Age

the age of the patient that is between 30 and 65 years.

See the Appendix for the DATA step program that uses the RANUNI function to randomly generate the data set values.

Using PROC MEANS

The MEANS procedure only requires a PROC statement to summarize all the numeric data in a data set. If you want to analyze specific variables, you must use the VAR statement. Unless you specify the FW= option or the MAXDEC= option in the PROC statement, the statistics use the default format.

The following program produces a report for the analysis variables Height, Weight, and Age. The statistics are shown using a field width of six and two decimal places.

```
ods html body='drive:\filename.htm';

proc means data=clinic_study fw=6 maxdec=2;
  var height weight age;
  title1 'Default Statistics';
run;
```

The HTML output that ODS produces for this example follows.

Default Statistics						
The MEANS Procedure						
Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Height	Patient height in inches	360	66.96	4.45	58.00	75.90
Weight	Patient weight in pounds	360	185.78	45.70	106.30	279.90
Age	Patient age	360	46.22	10.14	30.00	64.00

To further customize the report, you can use various options in the PROC statement. To generate statistics for each treatment and gender, you can include a CLASS statement in the PROC step. The following program produces such a report, while requesting specific statistics.

```
ods html body='drive:\filename.htm';

proc means data=clinic_study fw=6 maxdec=2
  nonobs n mean std median;
  var height weight age;
  class treatment gender;
  format treatment treatfmt. gender gendfmt.;
  title1 'Data Summary For Treatment and Gender';
run;
```

The report now contains four statistics: the number of observations, the mean, the standard deviation, and the median. The NONOBS option is used to suppress the column that displays the total number of observations for each unique combination of both class variable values. The FORMAT statement is used to assign user-defined formats to the two class variables. PROC MEANS uses the formats to create row labels for Treatment and Gender.

The HTML output that ODS produces follows.

Data Summary For Treatment and Gender							
The MEANS Procedure							
Treatment group	Gender	Variable	Label	N	Mean	Std Dev	Median
Placebo	Male	Height	Patient height in inches	44	69.51	3.37	69.25
		Weight	Patient weight in pounds	44	218.95	36.33	218.10
		Age	Patient age	44	45.80	11.16	44.00
	Female	Height	Patient height in inches	44	63.78	3.54	64.25
		Weight	Patient weight in pounds	44	155.46	30.97	157.90
		Age	Patient age	44	48.23	9.89	50.00
Drug	Male	Height	Patient height in inches	44	70.30	3.53	70.80
		Weight	Patient weight in pounds	44	215.13	31.35	216.00
		Age	Patient age	44	45.07	10.08	43.00
	Female	Height	Patient height in inches	57	64.05	3.53	64.00
		Weight	Patient weight in pounds	57	155.11	30.35	158.60
		Age	Patient age	57	46.14	9.82	45.00
Diet	Male	Height	Patient height in inches	43	70.01	3.19	69.60
		Weight	Patient weight in pounds	43	225.07	35.63	231.20
		Age	Patient age	43	46.58	8.94	47.00
	Female	Height	Patient height in inches	39	63.98	3.20	64.00
		Weight	Patient weight in pounds	39	150.26	29.66	147.20
		Age	Patient age	39	46.36	10.73	47.00
Drug and Diet	Male	Height	Patient height in inches	46	69.64	3.39	69.45
		Weight	Patient weight in pounds	46	216.62	31.64	214.95
		Age	Patient age	46	44.26	9.93	43.00
	Female	Height	Patient height in inches	43	64.86	3.59	65.00
		Weight	Patient weight in pounds	43	153.47	31.46	157.30
		Age	Patient age	43	47.47	10.80	46.00

Using PROC UNIVARIATE

The UNIVARIATE procedure also requires just a PROC statement to summarize all the numeric data in a data set. If you want to analyze specific variables, you must use the VAR statement. By default, the tests for location examine the hypothesis that the mean is equal to zero. However, this may not make sense for the data you are analyzing. Use the MU0= option in the PROC statement to test the hypothesis that the mean is equal to a specified value μ_0 .

To save space, the following program produces an extensive report for only one analysis variable Weight. You can perform a similar analysis for Height and Age by including these variables in the VAR statement.

```
ods html body='drive:\filename.htm';

proc univariate data=clinic_study nextrobs=3;
  var weight;
  id patient;
run;
```

The NEXTROBS= option is included in the PROC statement to control the number of extreme values shown in the report. The ID statement requests that a patient number identify the three highest and lowest extreme values.

The HTML output that ODS produces for this example follows.

The UNIVARIATE Procedure
Variable: Weight (Patient weight in pounds)

Moments			
N	360	Sum Weights	360
Mean	185.785	Sum Observations	66882.6
Std Deviation	45.7037365	Variance	2088.83153
Skewness	0.13991891	Kurtosis	-0.8032487
Uncorrected SS	13175674.4	Corrected SS	749890.519
Coeff Variation	24.6003372	Std Error Mean	2.40879841

Basic Statistical Measures			
Location		Variability	
Mean	185.7850	Std Deviation	45.70374
Median	183.0000	Variance	2089
Mode	171.2000	Range	173.60000
		Interquartile Range	64.20000

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

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	77.12767	Pr > t 	<.0001
Sign	M	180	Pr >= M 	<.0001
Signed Rank	S	32490	Pr >= S 	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	279.90
99%	278.60
95%	262.50
90%	253.35
75% Q3	217.40
50% Median	183.00
25% Q1	153.20
10%	122.65
5%	111.75
1%	107.80
0% Min	106.30

Extreme Observations					
Lowest			Highest		
Value	Patient	Obs	Value	Patient	Obs
106.3	F150	150	278.6	M251	251
106.5	F222	222	279.4	M250	250
106.7	F032	32	279.9	M354	354

Notice the message after the table of Basic Statistical Measures. The note informs you of multiple modes in the data. However, the procedure reports a single mode, the lowest value that occurs most often. To list all possible modes, use the MODES option in the PROC UNIVARIATE statement.

Using PROC TABULATE

The TABULATE procedure requires a PROC statement, either a CLASS statement or VAR statement, and a TABLE statement to display summary statistics in tabular form. The CLASS statement specifies any variables that you may use to group the data. The VAR statement identifies any analysis variables you want in the table. The TABLE statement provides instructions on how to construct the table.

To construct one-, two-, or three-dimensional tables, you specify a series of expressions. Dimension expressions are composed of elements and operators and are separated by commas. The structure of the table depends on the variables, keywords, and operators you use in the expressions.

When you specify a TABLE statement with three dimensions, the leftmost expression defines pages, the next expression defines rows, and the rightmost expression defines columns. For two dimensions, the left expression defines rows and the right expression defines columns. For a one-dimensional table, the expression defines columns.

The following program produces a one-dimensional table for the analysis variables Height, Weight, and Age. The TABLE statement does not request any statistics. Therefore, PROC TABULATE computes a default statistic, the sum, for each analysis variable. The program also includes a TITLE statement because PROC TABULATE does not automatically provide a title for the table.

```
ods html body='drive:\filename.htm';

title1 'Proc Tabulate: Default Table';
proc tabulate data=clinic_study;
  var age weight height;
  table height weight age;

run;
```

The table contains three columns or cells. The header for each column is the variable label and the statistic name, Sum. The HTML output that ODS produces follows.

Proc Tabulate: Default Table

Patient height in inches	Patient weight in pounds	Patient age
Sum	Sum	Sum
24106.60	66882.60	16638.00

The next program modifies the TABLE statement to request five statistics, the mean, standard deviation, minimum, maximum, and median. The asterisk, also called the nesting or crossing operator, tells PROC TABULATE to calculate statistics for the variable that occurs before it.

```
ods html body='drive:\filename.htm';

title1 'Proc Tabulate: Summary Statistics';
proc tabulate data=clinic_study;
  var age weight height;
  table (height weight age)
        *(mean std min max median);
run;
```

Because an asterisk is between a list of variables and a list of statistical keywords, the table contains 15 columns with the five statistics for each analysis variable. The HTML output that ODS produces for this example follows.

Proc Tabulate: Summary Statistics

Patient height in inches					Patient weight in pounds					Patient age				
Mean	Std	Min	Max	Median	Mean	Std	Min	Max	Median	Mean	Std	Min	Max	Median
66.96	4.45	58.00	75.90	67.10	185.78	45.70	106.30	279.90	183.00	46.22	10.14	30.00	64.00	45.00

This table is rather simple and uses the default format for the data cells. In contrast to the PROC MEANS report, it is harder to make quick comparisons because the statistics are spread across one wide row.

The next program computes the same statistics but constructs a two-dimensional table and takes advantage of the formatting and labeling capabilities in PROC TABULATE. The TABLES statement now includes a row expression and column expression.

```
ods html body='drive:\filename.htm';

title1 'Proc Tabulate: Customized Table';
proc tabulate data=clinic_study format=6.2;
  var age weight height;
  table height*(mean std min max median)
        weight*(mean std min max median)
        age*(mean*f=5.1 std ((min max)*f=4.0)
              median), all;
  keylabel std='Std Dev' min='Minimum'
           max='Maximum' all='';
run;
```

The column dimension uses the keyword ALL to create one column with *all* the statistics. The row dimension specifies the analysis variables and the statistics. The statistics for Age will use a different format for the mean, minimum, and maximum. The KEYLABEL statement provides new row labels for the statistical keywords STD, MIN, and MAX, and suppresses the column label for ALL.

The HTML output that ODS produces for this example follows.

Proc Tabulate: Customized Table

Patient height in inches	Mean	66.96
	Std Dev	4.45
	Minimum	58.00
	Maximum	75.90
	Median	67.10
Patient weight in pounds	Mean	185.78
	Std Dev	45.70
	Minimum	106.30
	Maximum	279.90
	Median	183.00
Patient age	Mean	46.2
	Std Dev	10.14
	Minimum	30
	Maximum	64
	Median	45.00

This table contains 15 rows with the five statistics for each analysis variable. However, it is not as concise as the PROC MEANS report. The next program shows how the order of the expressions changes the orientation of the table. By using the ALL keyword nested with the statistic keywords, you create a two-dimensional table that is three rows and five columns.

```
ods html body='drive:\filename.htm';

title1 'Proc Tabulate: Customized Table';
proc tabulate data=clinic_study format=6.2
  qmethod=p2;
  var age weight height;
  keylabel std='Std Dev' min='Minimum'
           max='Maximum' all='';
  table height weight age,
        all=' *(mean std min max median);
run;
```

The PROC statement also includes the QMETHOD=P2 option to change how PROC TABULATE computes the median. For a larger data set, this can reduce processing time. Notice however, the median is not the same value that PROC UNIVARIATE reports.

Proc Tabulate: Customized Table

	Mean	Std Dev	Minimum	Maximum	Median
Patient height in inches	66.96	4.45	58.00	75.90	67.17
Patient weight in pounds	185.78	45.70	106.30	279.90	183.71
Patient age	46.22	10.14	30.00	64.00	45.28

The final PROC TABULATE program adds a CLASS statement to construct a detailed tabular report that groups the observations by treatment and gender.

```
ods html body='drive:\filename.htm';

title1 'Proc Tabulate: Customized Table';
proc tabulate data=clinic_study format=6.2
    qmethod=p2;
    class treatment gender;
    var height weight age;
    table (age weight height)*(mean
        std='Std Dev' min max median),
        treatment='Treatment Group'*
        gender all='Total';
    format treatment treatfmt.
        gender gendfmt.;
run;
```

The ALL keyword in the row expression will summarize *all* of the categories for class variables across the columns. The KEYLABEL statement is omitted because labels are specified in the TABLE statement.

The table contains again contains 15 rows with five statistics for each analysis variable. The order of the variables in the report corresponds to the order they were listed in the row expression. The number of the columns in the table is nine. The last row is the statistic PROC TABULATE computes for the total of all the observations in a row. The HTML output that ODS produces for this example follows.

Proc Tabulate: Customized Table

		Treatment Group								Total
		Placebo		Drug		Diet		Drug and Diet		
		Gender		Gender		Gender		Gender		
		Male	Female	Male	Female	Male	Female	Male	Female	
Patient age	Mean	45.80	48.23	45.07	46.14	46.58	46.36	44.26	47.47	46.22
	Std Dev	11.16	9.89	10.08	9.82	8.94	10.73	9.93	10.80	10.14
	Min	30.00	30.00	30.00	30.00	30.00	30.00	30.00	30.00	30.00
	Max	64.00	63.00	64.00	63.00	63.00	64.00	63.00	64.00	64.00
	Median	45.17	49.19	41.98	45.96	46.91	47.77	41.97	46.05	46.20
Patient weight in pounds	Mean	218.95	155.46	215.13	155.11	225.07	150.26	216.62	153.47	185.79
	Std Dev	36.33	30.97	31.35	30.35	35.63	29.66	31.64	31.46	45.70
	Min	160.20	106.30	161.00	106.50	164.80	108.10	165.30	108.50	106.30
	Max	278.50	200.70	275.60	206.20	279.90	207.10	267.60	208.90	279.90
	Median	221.12	153.43	218.94	157.13	229.10	145.48	217.48	151.32	200.81
Patient height in inches	Mean	69.51	63.78	70.30	64.05	70.01	63.98	69.64	64.86	66.96
	Std Dev	3.37	3.54	3.53	3.53	3.19	3.20	3.39	3.59	4.45
	Min	64.30	58.10	64.20	58.00	64.90	58.50	64.00	58.10	58.00
	Max	75.90	69.80	75.90	69.90	75.60	68.60	75.90	69.90	75.90
	Median	69.58	63.88	70.99	63.79	69.21	63.82	69.93	64.95	67.97

PROC TABULATE provides several other statements and options to customize your reports. For additional instructions on how to use PROC TABULATE to construct sophisticated tabular reports, see *PROC TABULATE By Example* (Haworth, 1999) and the SUGI article by Kelley and McNeill (1999) .

PRODUCING A GRAPHICAL SUMMARY

The following program use two new graphics statements in the UNIVARIATE procedure to create and label a high-resolution frequency histogram for the analysis variable Weight.

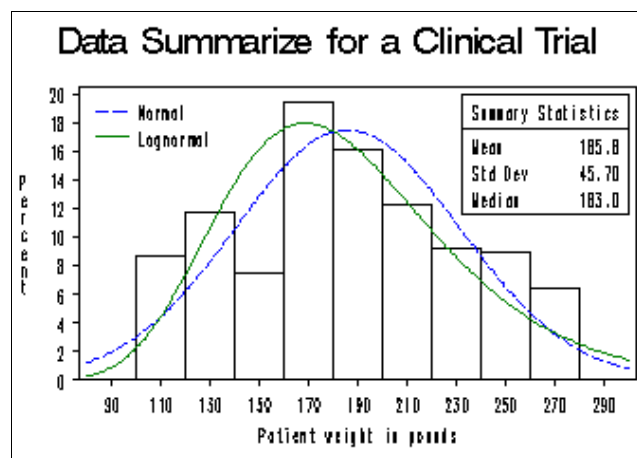
```
title1 'Data Summarize for a Clinical Trial';
proc univariate data=clinic_study noprint;
    var weight;
    histogram / normal(noprint color=blue l=3)
        lognormal(noprint color=green)
        vaxis=0 to 20 by 2
        midpoints=90 to 290 by 20;
    inset mean std='Std Dev' (6.2) median
        / position=ne format=5.1
        header='Summary Statistics';
    inset normal lognormal / position=nw
        noframe;
run;
```

Because the NOPRINT option is used in the PROC statement, PROC UNIVARIATE suppresses the default tables of statistics. The HISTOGRAM statement invokes graphics and produces a histogram and two fitted density curves. When a variable is omitted from the HISTOGRAM statement, PROC UNIVARIATE creates a histogram for all the variables listed in the VAR statement, which is only Weight. The NOPRINT options in the HISTOGRAM statement suppress any tables of statistics that summarize the fitted density curves.

Other options in the histogram statement control the graph's appearance. The NORMAL option superimposes the fitted density curve for a normal distribution. The LOGNORMAL option superimposes the fitted density curve for the lognormal distribution. The COLOR= option specifies the line colors for the fitted density curves. 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. The VAXIS= option specifies the tick mark labels for the vertical axis.

The INSET statements that follow the HISTOGRAM statement inset two tables directly in the graph. The statistical keywords in the statement tell PROC UNIVARIATE which values to include in the table. STD= requests a customized label and a field width of six and two decimal places. The keywords NORMAL and LOGNORMAL alone inset a colored line and the distribution name that you can use as a key for the density curves. The FORMAT= option specifies the default format for the statistics in the table. The POSITION= option controls the placement of the insets in the graph. The HEADER= option provides a label for the table and the NOFRAME option suppresses the frame around the second inset table.

A black and white image of graph that the program produces follows. The key for the density curves is located in the northwest corner of the graph while the table of statistics is in the northeast corner.



The next program examines the variable Weight again. This time a CLASS statement is added to the PROC step to create a comparative histogram for treatment and gender.

```

Title1 'Data Summarize for a Clinical Trial';
proc univariate data=clinic_study noprint;
  var weight;
  class gender treatment;
  histogram / vscale=count
              midpoints=100 to 300 by 20
              nrow=2 ncol=4 intertile=1
              vaxis=0 to 20 by 2;
  inset n='No Obs' (2.) mean
        std='Std Dev' (6.2) median /
        height=1 font=swissb position=ne
        header='Summary Statistics'
        format=5.1;
  format treatment treatfmt.
        Gender gendfmt. ;
run;

```

A two-way comparative histogram is created for the analysis variable. PROC UNIVARIATE produces a component histogram for each level (distinct combination of values) of the class variables. The NROWS= and NCOLS= options request a 2 x 4 arrangement for the component histograms. By default, the arrangement for the component histograms is 2 x 2. The INTERTILE= inserts a space of one percent screen unit between them. The VSCALE= option requests the vertical axis scale in units of the number of observations per data unit. The FORMAT statement assigns user-defined formats to the class variables Treatment and Gender so that PROC UNIVARIATE uses formatted values to label each component histogram.

The INSET statement insets a table directly on each component histogram with the number of observations, the mean, the standard deviation, and median. A customized format and label are assigned to the keywords N and STD. The FORMAT= option assigns a field width of five and one decimal place to the other statistics in the inset table. The HEIGHT= and FONT= options request a specific height and font for the text. The POSITION= option controls the

placement of the inset in each component histogram.

Figure 1 in the Appendix contains a black and white image of the comparative histogram. The graphic allows you quickly summarize data and make comparisons by gender and treatment.

PROC UNIVARIATE provides various other options to customize high-resolution histograms. You can also use the PROBPLOT statement and QQPLOT statement to generate high-resolution plots that allow you examine the underlying distribution of your data. See the SAS *Procedures Guide, Version 8* for more information.

CONCLUSION

The reporting procedures in Base SAS are very powerful and provide you with a variety of ways to summarize your data. You have seen how quickly you can use PROC MEANS or PROC UNIVARIATE to produce a report. With some work and creativity you can use PROC TABULATE to design sophisticated tabular reports. For a graphical analysis of the underlying distribution of your data, you can use new statements in PROC UNIVARIATE. For a more complete examination of the data graphically, consider using SAS/INSIGHT® software.

REFERENCES

- Haworth, L. (1999), *PROC TABULATE By Example*, Cary, NC: SAS Institute Inc.
- Jain, R. and Chlamtac, I. (1985), "The P² Algorithm for Dynamic Calculation of Quantiles and Histograms Without Sorting Observations," *Communications of the Association of Computing Machinery*, 28:10.
- Kelley, D. and McNeill, S. (1999), "Getting Stylish with Version 7 Base Reporting," *Proceedings of the Twenty-Fourth Annual SAS Users Group International Conference*.
- SAS Institute Inc. (1999), *SAS Procedures Guide*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1999), *The Complete Guide to the SAS Output Delivery System*, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

SAS Institute Inc.
 SAS Campus Drive
 Cary, NC 27513
 Work Phone: (919) 677-8001 x6429
 Fax: (919) 677-4444
 Email: Lynne.Bresler@sas.com

TRADEMARKS

SAS and SAS/GRAPH are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Appendix A

Below is the source code to format and generate the CLINIC_STUDY data set.

```
Proc format;
  value gendfmt    0='Male'
                  1='Female';
  value treatfmt   0='Placebo'
                  1='Drug'
                  2='Diet'
                  3='Drug and Diet';
run;

data clinic_study;
  drop I;
  label Patient='Patient Number'
        Height='Patient height in inches'
        Weight='Patient weight in pounds'
        Treatment='Treatment group'
        Age='Patient age' ;
  do I=1 to 360;
    Gender=(ranuni(700)<=0.5);
    if gender=0 then do; /* male patient */
      Patient='M'||(put(I,z3.));
      Height=round((ranuni(22878)*12+64),.1);
      Weight=round((ranuni(2179)*120+160),.1);
    end;
    else do; /* female patient */
      Patient='F'||(put(I,z3.));
      Height=round((ranuni(22878)*12+58),.1);
      Weight=round((ranuni(2179)*105+105),.1);
    end;
    Age=int(ranuni(51602)*35)+30;
    Treatment=int(ranuni(36830)*4)+0;
    output;
  end;
run;
```

Figure 1: A Comparative Histogram From an Univariate Analysis

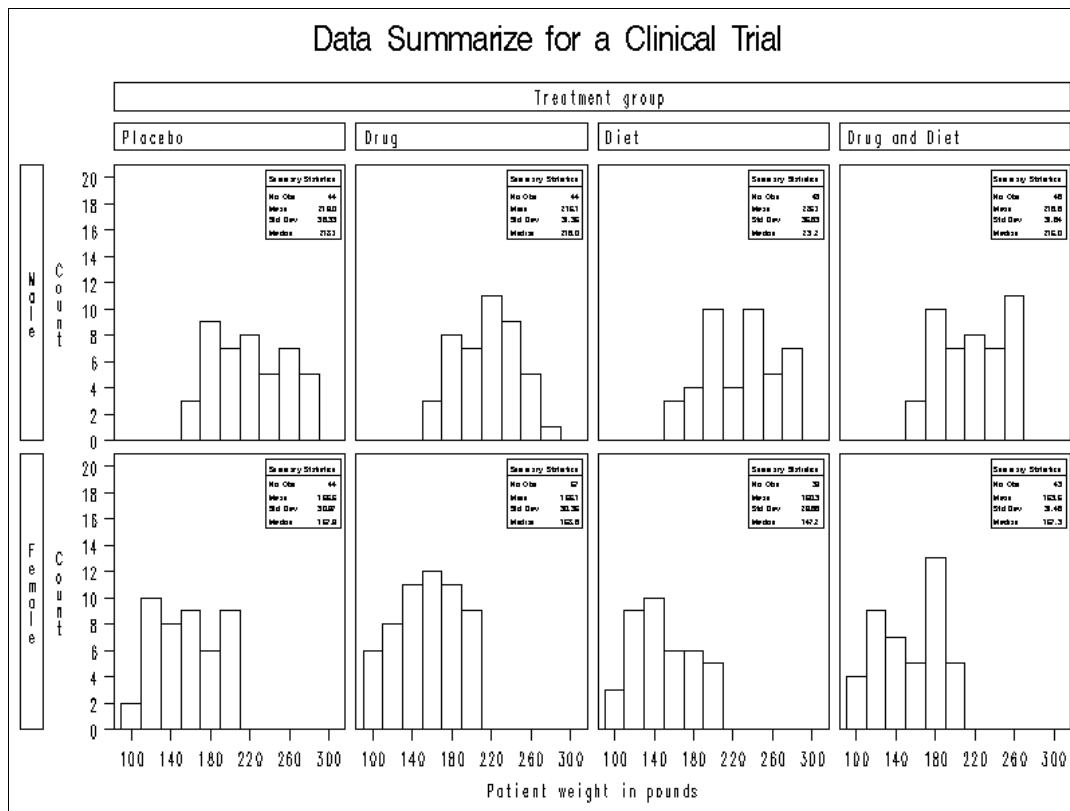


Table 1: Common Summary Statistics

Statistic	MEANS/ SUMMARY	UNIVARIATE	TABULATE	REPORT	CORR	SQL
Number of missing values	X	X	X	X		X
Number of nonmissing values	X	X	X	X	X	X
Number of observations	X	X				X
Sum of weights	X	X	X	X	X	X
Mean	X	X	X	X	X	X
Sum	X	X	X	X	X	X
Extreme values	X	X				
Minimum	X	X	X	X	X	X
Maximum	X	X	X	X	X	X
Range	X	X	X	X		X
Uncorrected sum of squares	X	X	X	X	X	X
Corrected sum of squares	X	X	X	X	X	X
Variance	X	X	X	X	X	X
Covariance					X	
Standard deviation	X	X	X	X	X	X
Standard error of the mean	X	X	X	X		X
Coefficient of variation	X	X	X	X		X
Skewness	X	X	X			
Kurtosis	X	X	X			
Confidence Limits the mean	X	X				
the variance		X				
quantiles		X				
Median	X	X	X		X	
Mode		X				
Percentiles/Deciles/ Quartiles	X	X	X			
Student's t test mean=0	X	X	X	X		X
mean= μ		X				
Nonparametric tests for location		X				
Tests for normality		X				
Correlation coefficients					X	