Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.
## Contents

<table>
<thead>
<tr>
<th>Chapter 1.</th>
<th>What’s New in the Base SAS 9.4 Statistical Procedures</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2.</td>
<td>The CORR Procedure</td>
<td>3</td>
</tr>
<tr>
<td>Chapter 3.</td>
<td>The FREQ Procedure</td>
<td>69</td>
</tr>
<tr>
<td>Chapter 4.</td>
<td>The UNIVARIATE Procedure</td>
<td>263</td>
</tr>
</tbody>
</table>

Subject Index 525

Syntax Index 535
Chapter 1
What’s New in the Base SAS 9.4 Statistical Procedures

Overview
The first maintenance release of Base SAS 9.4 includes enhancements to the FREQ and UNIVARIATE statistical procedures.

FREQ Procedure Enhancements

• The new OR(CL=SCORE) option provides score confidence limits for the odds ratio. The RISKDIFF(COMMON) option provides Mantel-Haenszel and summary score estimates of the common risk (proportion) difference. The score confidence limits can be displayed in the odds ratio plot, and the common risk difference can be displayed in the risk difference plot.

• The MIDP option in the EXACT statement produces mid $p$-values for exact tests.

• The new COLORSTAT= option for mosaic plots colors the tiles according to the values of the Pearson residuals or the standardized residuals. The SCALE=GROUPPERCENT option for two-way frequency plots displays the row or column percentages (instead of the overall percentages). The CLDISPLAY=SERIFARROW and CLDISPLAY=LINEARROW options are now available to control the error bars in odds ratio, relative risk, risk difference, and kappa plots.

• The CROSSLIST(PEARSONRES) option displays the Pearson residuals in the CROSSLIST table.

UNIVARIATE Procedure Enhancements

• The HISTOGRAM statement now supports the OVERLAY option, which overlays histograms that are associated with different levels of a CLASS variable onto a single plot.

• Maximum likelihood estimation for Johnson $S_U$ distribution parameters has been improved. You can fit a Johnson $S_U$ distribution by specifying the SU option in the HISTOGRAM statement. You can request maximum likelihood estimation by specifying the FITMETHOD=MLE suboption, and you can control the optimizer that performs the estimation by specifying the new OPTBOUNDDRANGE=, OPTMAXITER=, OPTMAXSTARTS=, OPTPRINT, OPTSEED=, and OPTTOLERANCE= suboptions.
Chapter 2
The CORR Procedure

Contents

Overview: CORR Procedure ................................................. 4
Getting Started: CORR Procedure ........................................ 5
Syntax: CORR Procedure .................................................... 9
  PROC CORR Statement .................................................. 9
  BY Statement .......................................................... 17
  FREQ Statement ...................................................... 17
  ID Statement .......................................................... 18
  PARTIAL Statement ................................................... 18
  VAR Statement ......................................................... 18
  WEIGHT Statement ................................................... 18
  WITH Statement ...................................................... 19
Details: CORR Procedure .................................................... 19
  Pearson Product-Moment Correlation ............................. 19
  Spearman Rank-Order Correlation ................................. 21
  Kendall’s Tau-b Correlation Coefficient ......................... 22
  Hoeffding Dependence Coefficient ............................... 23
  Partial Correlation .................................................... 23
  Fisher’s z Transformation .......................................... 25
  Polychoric Correlation .............................................. 28
  Polyserial Correlation .............................................. 29
  Cronbach’s Coefficient Alpha ...................................... 31
  Confidence and Prediction Ellipses .............................. 32
  Missing Values ....................................................... 33
  In-Database Computation ........................................... 34
  Output Tables ......................................................... 34
  Output Data Sets ..................................................... 35
  ODS Table Names .................................................... 37
  ODS Graphics ........................................................ 38
Examples: CORR Procedure .................................................. 38
  Example 2.1: Computing Four Measures of Association .... 38
  Example 2.2: Computing Correlations between Two Sets of Variables 43
  Example 2.3: Analysis Using Fisher’s z Transformation .... 48
  Example 2.4: Applications of Fisher’s z Transformation .... 50
  Example 2.5: Computing Polyserial Correlations ............... 54
  Example 2.6: Computing Cronbach’s Coefficient Alpha ...... 56
  Example 2.7: Saving Correlations in an Output Data Set .... 59
Overview: CORR Procedure

The CORR procedure computes Pearson correlation coefficients, three nonparametric measures of association, polyserial correlation coefficients, and the probabilities associated with these statistics. The correlation statistics include the following:

- Pearson product-moment correlation
- Spearman rank-order correlation
- Kendall’s tau-b coefficient
- Hoeffding’s measure of dependence, $D$
- Pearson, Spearman, and Kendall partial correlation
- polychoric correlation
- polyserial correlation

Pearson product-moment correlation is a parametric measure of a linear relationship between two variables. For nonparametric measures of association, Spearman rank-order correlation uses the ranks of the data values and Kendall’s tau-b uses the number of concordances and discordances in paired observations. Hoeffding’s measure of dependence is another nonparametric measure of association that detects more general departures from independence. A partial correlation provides a measure of the correlation between two variables after controlling the effects of other variables.

Polyserial correlation measures the correlation between two continuous variables with a bivariate normal distribution, where only one variable is observed directly. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values.

A related type of correlation, polychoric correlation, measures the correlation between two unobserved variables with a bivariate normal distribution. Information about these variables is obtained through two corresponding observed ordinal variables that are derived from the unobserved variables by classifying their values into finite sets of discrete, ordered values.

When only one set of analysis variables is specified, the default correlation analysis includes descriptive statistics for each analysis variable and pairwise Pearson correlation statistics for these variables. You can also compute Cronbach’s coefficient alpha for estimating reliability.

When two sets of analysis variables are specified, the default correlation analysis includes descriptive statistics for each analysis variable and pairwise Pearson correlation statistics between the two sets of variables.
For a Pearson or Spearman correlation, the Fisher’s $z$ transformation can be used to derive its confidence limits and a $p$-value under a specified null hypothesis $H_0: \rho = \rho_0$. Either a one-sided or a two-sided alternative is used for these statistics.

When the relationship between two variables is nonlinear or when outliers are present, the correlation coefficient might incorrectly estimate the strength of the relationship. Plotting the data enables you to verify the linear relationship and to identify the potential outliers. If ODS Graphics is enabled, scatter plots and a scatter plot matrix can be created via the Output Delivery System (ODS). Confidence and prediction ellipses can also be added to the scatter plot. See the section “Confidence and Prediction Ellipses” on page 32 for a detailed description of the ellipses.

You can save the correlation statistics in a SAS data set for use with other statistical and reporting procedures.

---

**Getting Started: CORR Procedure**

The following statements create the data set `Fitness`, which has been altered to contain some missing values:

```sas
*----------------- Data on Physical Fitness -----------------*  
| These measurements were made on men involved in a physical |  
| fitness course at N.C. State University. |  
| The variables are Age (years), Weight (kg), |  
| Runtime (time to run 1.5 miles in minutes), and |  
| Oxygen (oxygen intake, ml per kg body weight per minute) |  
| Certain values were changed to missing for the analysis. |  
*------------------------------------------------------------*;  
data Fitness;  
    input Age Weight Oxygen RunTime @@;  
datalines;  
44 89.47 44.609 11.37   40 75.07 45.313 10.07  
44 85.84 54.297 8.65  42 68.15 59.571 8.17  
38 89.02 49.874 .    47 77.45 44.811 11.63  
40 75.98 45.681 11.95  43 81.19 49.091 10.85  
45 81.42 39.442 13.08 38 81.87 60.055 8.63  
44 73.03 50.541 10.13 45 87.66 37.388 14.03  
45 66.45 44.754 11.12 47 79.15 47.273 10.60  
54 83.12 51.855 10.33 49 81.42 49.156 8.95  
51 69.63 40.836 10.95 51 77.91 46.672 10.00  
48 91.63 46.774 10.25 49 73.37 .  10.08  
57 73.37 39.407 12.63 54 79.38 46.080 11.17  
52 76.32 45.441 9.63  50 70.87 54.625 8.92  
51 67.25 45.118 11.08 54 91.63 39.203 12.88  
51 73.71 45.790 10.47 57 59.08 50.545 9.93  
49 76.32 . .     48 61.24 47.920 11.50  
52 82.78 47.467 10.50  
;  
```
The following statements invoke the CORR procedure and request a correlation analysis:

```plaintext
ods graphics on;
proc corr data=Fitness plots=matrix(histogram);
run;
ods graphics off;
```

The “Simple Statistics” table in Figure 2.1 displays univariate statistics for the analysis variables.

**Figure 2.1** Univariate Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>31</td>
<td>47.67742</td>
<td>5.21144</td>
<td>1478</td>
<td>38.00000</td>
<td>57.00000</td>
</tr>
<tr>
<td>Weight</td>
<td>31</td>
<td>77.44452</td>
<td>8.32857</td>
<td>2401</td>
<td>59.08000</td>
<td>91.63000</td>
</tr>
<tr>
<td>Oxygen</td>
<td>29</td>
<td>47.22721</td>
<td>5.47718</td>
<td>1370</td>
<td>37.38800</td>
<td>60.05500</td>
</tr>
<tr>
<td>RunTime</td>
<td>29</td>
<td>10.67414</td>
<td>1.39194</td>
<td>309.5500</td>
<td>8.17000</td>
<td>14.03000</td>
</tr>
</tbody>
</table>

By default, all numeric variables not listed in other statements are used in the analysis. Observations with nonmissing values for each variable are used to derive the univariate statistics for that variable.

The “Pearson Correlation Coefficients” table in Figure 2.2 displays the Pearson correlation, the $p$-value under the null hypothesis of zero correlation, and the number of nonmissing observations for each pair of variables.
By default, Pearson correlation statistics are computed from observations with nonmissing values for each pair of analysis variables. Figure 2.2 displays a correlation of $-0.86843$ between Runtime and Oxygen, which is significant with a $p$-value less than 0.0001. That is, there exists an inverse linear relationship between these two variables. As Runtime (time to run 1.5 miles in minutes) increases, Oxygen (oxygen intake, ml per kg body weight per minute) decreases.
When you use the PLOTS=MATRIX(HISTOGRAM) option, the CORR procedure displays a symmetric matrix plot for the analysis variables in Figure 2.3. The histograms for these analysis variables are also displayed on the diagonal of the matrix plot. This inverse linear relationship between the two variables, Oxygen and Runtime, is also shown in the plot.

Note that ODS Graphics must be enabled and you must specify the PLOTS= option to produce graphs. For more information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS” (SAS/STAT User’s Guide).

**Figure 2.3 Symmetric Matrix Plot**
Syntax: CORR Procedure

The following statements are available in PROC CORR:

```
PROC CORR < options > ;
    BY variables ;
    FREQ variable ;
    ID variables ;
    PARTIAL variables ;
    VAR variables ;
    WEIGHT variable ;
    WITH variables ;
```

The BY statement specifies groups in which separate correlation analyses are performed.

The FREQ statement specifies the variable that represents the frequency of occurrence for other values in the observation.

The ID statement specifies one or more additional tip variables to identify observations in scatter plots and scatter plot matrices.

The PARTIAL statement identifies controlling variables to compute Pearson, Spearman, or Kendall partial-correlation coefficients.

The VAR statement lists the numeric variables to be analyzed and their order in the correlation matrix. If you omit the VAR statement, all numeric variables not listed in other statements are used.

The WEIGHT statement identifies the variable whose values weight each observation to compute Pearson product-moment correlation.

The WITH statement lists the numeric variables with which correlations are to be computed.

The PROC CORR statement is the only required statement for the CORR procedure. The rest of this section provides detailed syntax information for each of these statements, beginning with the PROC CORR statement. The remaining statements are presented in alphabetical order.

PROC CORR Statement

```
PROC CORR < options > ;
```

Table 2.1 summarizes the options available in the PROC CORR statement.

```
Table 2.1    Summary of PROC CORR Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Sets</strong></td>
<td></td>
</tr>
<tr>
<td>DATA=</td>
<td>Specifies the input data set</td>
</tr>
<tr>
<td>OUTH=</td>
<td>Specifies the output data set with Hoeffding’s $D$ statistics</td>
</tr>
<tr>
<td>OUTK=</td>
<td>Specifies the output data set with Kendall correlation statistics</td>
</tr>
<tr>
<td>OUTP=</td>
<td>Specifies the output data set with Pearson correlation statistics</td>
</tr>
</tbody>
</table>
```
Table 2.1 continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUTS=</td>
<td>Specifies the output data set with Spearman correlation statistics</td>
</tr>
<tr>
<td><strong>Statistical Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>EXCLNPWGT</td>
<td>Excludes observations with nonpositive weight values from the analysis</td>
</tr>
<tr>
<td>FISHER</td>
<td>Requests correlation statistics using Fisher’s z transformation</td>
</tr>
<tr>
<td>HOEFFDING</td>
<td>Requests Hoeffding’s measure of dependence, D</td>
</tr>
<tr>
<td>KENDALL</td>
<td>Requests Kendall’s tau-b</td>
</tr>
<tr>
<td>NOMISS</td>
<td>Excludes observations with missing analysis values from the analysis</td>
</tr>
<tr>
<td>PEARSON</td>
<td>Requests Pearson product-moment correlation</td>
</tr>
<tr>
<td>POLYCHORIC</td>
<td>Requests polychoric correlation</td>
</tr>
<tr>
<td>POLYSERIAL</td>
<td>Requests polyserial correlation</td>
</tr>
<tr>
<td>SPEARMAN</td>
<td>Requests Spearman rank-order correlation</td>
</tr>
<tr>
<td><strong>Pearson Correlation Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA</td>
<td>Computes Cronbach’s coefficient alpha</td>
</tr>
<tr>
<td>COV</td>
<td>Computes covariances</td>
</tr>
<tr>
<td>CSSCP</td>
<td>Computes corrected sums of squares and crossproducts</td>
</tr>
<tr>
<td>FISHER</td>
<td>Computes correlation statistics based on Fisher’s z transformation</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Specifies the singularity criterion</td>
</tr>
<tr>
<td>SSCP</td>
<td>Computes sums of squares and crossproducts</td>
</tr>
<tr>
<td>VARDEF=</td>
<td>Specifies the divisor for variance calculations</td>
</tr>
<tr>
<td><strong>ODS Output Graphics</strong></td>
<td></td>
</tr>
<tr>
<td>PLOTS=MATRIX</td>
<td>Displays the scatter plot matrix</td>
</tr>
<tr>
<td>PLOTS=SCATTER</td>
<td>Displays scatter plots for pairs of variables</td>
</tr>
<tr>
<td><strong>Printed Output</strong></td>
<td></td>
</tr>
<tr>
<td>BEST=</td>
<td>Displays the specified number of ordered correlation coefficients</td>
</tr>
<tr>
<td>NOCORR</td>
<td>Suppresses Pearson correlations</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>Suppresses all printed output</td>
</tr>
<tr>
<td>NOPROB</td>
<td>Suppresses p-values</td>
</tr>
<tr>
<td>NOSIMPLE</td>
<td>Suppresses descriptive statistics</td>
</tr>
<tr>
<td>RANK</td>
<td>Displays ordered correlation coefficients</td>
</tr>
</tbody>
</table>

The following options can be used in the PROC CORR statement. They are listed in alphabetical order.

**ALPHA**

calculates and prints Cronbach’s coefficient alpha. PROC CORR computes separate coefficients using raw and standardized values (scaling the variables to a unit variance of 1). For each VAR statement variable, PROC CORR computes the correlation between the variable and the total of the remaining variables. It also computes Cronbach’s coefficient alpha by using only the remaining variables.

If a WITH statement is specified, the ALPHA option is invalid. When you specify the ALPHA option, the Pearson correlations will also be displayed. If you specify the OUTP= option, the output data set also contains observations with Cronbach’s coefficient alpha. If you use the PARTIAL statement, PROC CORR calculates Cronbach’s coefficient alpha for partialled variables. See the section “Partial Correlation” on page 23 for details.
**BEST=n**

prints the $n$ highest correlation coefficients for each variable, $n \geq 1$. Correlations are ordered from highest to lowest in absolute value. Otherwise, PROC CORR prints correlations in a rectangular table, using the variable names as row and column labels.

If you specify the HOEFFDING option, PROC CORR displays the $D$ statistics in order from highest to lowest.

**COV**

displays the variance and covariance matrix. When you specify the COV option, the Pearson correlations will also be displayed. If you specify the OUTP= option, the output data set also contains the covariance matrix with the corresponding _TYPE_ variable value ‘COV.’ If you use the PARTIAL statement, PROC CORR computes a partial covariance matrix.

**CSSCP**

displays a table of the corrected sums of squares and crossproducts. When you specify the CSSCP option, the Pearson correlations will also be displayed. If you specify the OUTP= option, the output data set also contains a CSSCP matrix with the corresponding _TYPE_ variable value ‘CSSCP.’ If you use a PARTIAL statement, PROC CORR prints both an unpartial and a partial CSSCP matrix, and the output data set contains a partial CSSCP matrix.

**DATA=SAS-data-set**

names the SAS data set to be analyzed by PROC CORR. By default, the procedure uses the most recently created SAS data set.

**EXCLNPWGT**

**EXCLNPWGTS**

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

**FISHER** <(fisher-options)> 

requests confidence limits and $p$-values under a specified null hypothesis, $H_0: \rho = \rho_0$, for correlation coefficients by using Fisher’s $z$ transformation. These correlations include the Pearson correlations and Spearman correlations.

The following fisher-options are available:

**ALPHA=** $\alpha$

specifies the level of the confidence limits for the correlation, $100(1 - \alpha)\%$. The value of the ALPHA= option must be between 0 and 1, and the default is ALPHA=0.05.

**BIASADJ=YES | NO**

specifies whether or not the bias adjustment is used in constructing confidence limits. The BIASADJ=YES option also produces a new correlation estimate that uses the bias adjustment. By default, BIASADJ=YES.

**RHO0=** $\rho_0$

specifies the value $\rho_0$ in the null hypothesis $H_0: \rho = \rho_0$, where $-1 < \rho_0 < 1$. By default, RHO0=0.
Chapter 2: The CORR Procedure

**TYPE=LOWER | UPPER | TWOSIDED**

specifies the type of confidence limits. The **TYPE=LOWER** option requests a lower confidence limit for the test of the one-sided hypothesis \( H_0: \rho \leq \rho_0 \) against the alternative hypothesis \( H_1: \rho > \rho_0 \), the **TYPE=UPPER** option requests an upper confidence limit for the test of the one-sided hypothesis \( H_0: \rho \geq \rho_0 \) against the alternative hypothesis \( H_1: \rho < \rho_0 \), and the default **TYPE=TWOSIDED** option requests two-sided confidence limits for the test of the hypothesis \( H_0: \rho = \rho_0 \).

**HOEFFDING**

requests a table of Hoeffding’s \( D \) statistics. This \( D \) statistic is 30 times larger than the usual definition and scales the range between −0.5 and 1 so that large positive values indicate dependence. The **HOEFFDING** option is invalid if a WEIGHT or PARTIAL statement is used.

**KENDALL**

requests a table of Kendall’s tau-b coefficients based on the number of concordant and discordant pairs of observations. Kendall’s tau-b ranges from −1 to 1.

The **KENDALL** option is invalid if a WEIGHT statement is used. If you use a PARTIAL statement, probability values for Kendall’s partial tau-b are not available.

**NOCORR**

suppresses displaying of Pearson correlations. If you specify the OUTP= option, the data set type remains CORR. To change the data set type to COV, CSSCP, or SSCP, use the **TYPE=** data set option.

**NOMISS**

excludes observations with missing values from the analysis. Otherwise, PROC CORR computes correlation statistics by using all of the nonmissing pairs of variables. Using the **NOMISS** option is computationally more efficient.

**NOPRINT**

suppresses all displayed output, which also includes output produced with ODS Graphics. Use the **NOPRINT** option if you want to create an output data set only.

**NOPROB**

suppresses displaying the probabilities associated with each correlation coefficient.

**NOSIMPLE**

suppresses printing simple descriptive statistics for each variable. However, if you request an output data set, the output data set still contains simple descriptive statistics for the variables.

**OUTH=** `output-data-set`

creates an output data set that contains Hoeffding’s \( D \) statistics. The contents of the output data set are similar to those of the OUTP= data set. When you specify the **OUTH=** option, the Hoeffding’s \( D \) statistics will be displayed.

**OUTK=** `output-data-set`

creates an output data set that contains Kendall correlation statistics. The contents of the output data set are similar to those of the OUTP= data set. When you specify the **OUTK=** option, the Kendall correlation statistics will be displayed.
**OUTP=** *output-data-set*

**OUT=** *output-data-set*

creates an output data set that contains Pearson correlation statistics. This data set also includes means, standard deviations, and the number of observations. The value of the _TYPE_ variable is ‘CORR.’ When you specify the OUTP= option, the Pearson correlations will also be displayed. If you specify the ALPHA option, the output data set also contains six observations with Cronbach’s coefficient alpha.

**OUTS=** *SAS-data-set*

creates an output data set that contains Spearman correlation coefficients. The contents of the output data set are similar to those of the OUTP= data set. When you specify the OUTS= option, the Spearman correlation coefficients will be displayed.

**PEARSON**

requests a table of Pearson product-moment correlations. The correlations range from −1 to 1. If you do not specify the HOEFFDING, KENDALL, SPEARMAN, POLYCHORIC, POLYSERIAL, OUTH=, OUTK=, or OUTS= option, the CORR procedure produces Pearson product-moment correlations by default. Otherwise, you must specify the PEARSON, ALPHA, COV, CSSCP, SSCP, or OUT= option for Pearson correlations. Also, if a scatter plot or a scatter plot matrix is requested, the Pearson correlations will be displayed.

**PLOTS < ( MAXPOINTS=NONE | \( n \) ) > = plot-request**

requests statistical graphics via the Output Delivery System (ODS).

ODS Graphics must be enabled before plots can be requested. For example:

```plaintext
ods graphics on;
proc corr data=Fitness plots=matrix(histogram);
run;
ods graphics off;
```


The global plot option MAXPOINTS= specifies that plots with elements that require processing more than \( n \) points be suppressed. The default is MAXPOINTS=5000. This limit is ignored if you specify MAXPOINTS=NONE. The plot request options include the following:

**ALL**

produces all appropriate plots.

**MATRIX < (matrix-options) >**

requests a scatter plot matrix for variables. That is, the procedure displays a symmetric matrix plot with variables in the VAR list if a WITH statement is not specified. Otherwise, the procedure displays a rectangular matrix plot with the WITH variables appearing down the side and the VAR variables appearing across the top.
NONE
suppresses all plots.

SCATTER < (scatter-options)>
requests scatter plots for pairs of variables. That is, the procedure displays a scatter plot for each applicable pair of distinct variables from the VAR list if a WITH statement is not specified. Otherwise, the procedure displays a scatter plot for each applicable pair of variables, one from the WITH list and the other from the VAR list.

When a scatter plot or a scatter plot matrix is requested, the Pearson correlations will also be displayed.

The available matrix-options are as follows:

HIST | HISTOGRAM
displays histograms of variables in the VAR list (specified in the VAR statement) in the symmetric matrix plot.

NVAR=ALL | n
specifies the maximum number of variables in the VAR list to be displayed in the matrix plot, where \( n > 0 \). The NVAR=ALL option uses all variables in the VAR list. By default, NVAR=5.

NWITH=ALL | n
specifies the maximum number of variables in the WITH list (specified in the WITH statement) to be displayed in the matrix plot, where \( n > 0 \). The NWITH=ALL option uses all variables in the WITH list. By default, NWITH=5.

If the resulting maximum number of variables in the VAR or WITH list is greater than 10, only the first 10 variables in the list are displayed in the scatter plot matrix.

The available scatter-options are as follows:

ALPHA=\( \alpha \)
specifies the \( \alpha \) values for the confidence or prediction ellipses to be displayed in the scatter plots, where \( 0 < \alpha < 1 \). For each \( \alpha \) value specified, a \((1-\alpha)\) confidence or prediction ellipse is created. By default, \( \alpha = 0.05 \).

ELLIPSE=PREDICTION | CONFIDENCE | NONE
requests prediction ellipses for new observations (ELLIPSE=PREDICTION), confidence ellipses for the mean (ELLIPSE=CONFIDENCE), or no ellipses (ELLIPSE=NONE) to be created in the scatter plots. By default, ELLIPSE=PREDICTION.

NOINSET
suppresses the default inset of summary information for the scatter plot. The inset table contains the number of observations (Observations) and correlation.

NVAR=ALL | n
specifies the maximum number of variables in the VAR list (specified in the VAR statement) to be displayed in the plots, where \( n > 0 \). The NVAR=ALL option uses all variables in the VAR list. By default, NVAR=5.
NWITH=ALL | $n$

specifies the maximum number of variables in the WITH list (specified in the WITH statement) to be displayed in the plots, where $n > 0$. The NWITH=ALL option uses all variables in the WITH list. By default, NWITH=5.

If the resulting maximum number of variables in the VAR or WITH list is greater than 10, only the first 10 variables in the list are displayed in the scatter plots.

POLYCHORIC <(options)> 

requests a table of polychoric correlation coefficients. A polychoric correlation measures the correlation between two unobserved, continuous variables that have a bivariate normal distribution. Information about each unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values. If you specify a WEIGHT statement, the POLYCHORIC option is not applicable.

You can specify the following options for computing polychoric correlation:

CONVERGE=$p$

specifies the convergence criterion. The value $p$ must be between 0 and 1. The iterations are considered to have converged when the absolute change in the parameter estimates between iteration steps is less than $p$ for each parameter—that is, for the correlation and the thresholds for the unobserved continuous variable that define the categories for the ordinal variable. By default, CONVERGE=0.0001.

MAXITER=number

specifies the maximum number of iterations. The iterations stop when the number of iterations exceeds number. By default, MAXITER=200.

NGROUPS=ALL | $n$

specifies the maximum number of groups allowed for each ordinal variable, where $n > 1$. NGROUPS=ALL allows an unlimited number of groups in each ordinal variable. Otherwise, if the number of groups exceeds the specified number $n$, polychoric correlations are not computed for the affected pairs of variables. By default, NGROUPS=20.

POLYSERIAL <(options)> 

requests a table of polyserial correlation coefficients. A polyserial correlation measures the correlation between two continuous variables with a bivariate normal distribution, where one variable is observed and the other is unobserved. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values. If you specify a WEIGHT statement, the POLYSERIAL option is not applicable.

You can specify the following options for computing polyserial correlation:

CONVERGE=$p$

specifies the convergence criterion. The value $p$ must be between 0 and 1. The iterations are considered to have converged when the absolute change in the parameter estimates between iteration steps is less than $p$ for each parameter—that is, for the correlation and the thresholds for the unobserved continuous variable that define the categories for the ordinal variable. By default, CONVERGE=0.0001.
Chapter 2: The CORR Procedure

**MAXITER=number**
specifies the maximum number of iterations. The iterations stop when the number of iterations exceeds number. By default, MAXITER=200.

**NGROUPS=ALL | n**
specifies the maximum number of groups allowed for each ordinal variable, where n > 1. NGROUPS=ALL allows an unlimited number of groups in each ordinal variable. Otherwise, if the number of groups exceeds the specified number n, polyserial correlations are not computed for the affected pairs of variables. By default, NGROUPS=20.

**ORDINAL=WITH | VAR**
specifies the ordinal variable list. The ORDINAL=WITH option specifies that the ordinal variables are provided in the WITH statement, and the continuous variables are provided in the VAR statement. The ORDINAL=VAR option specifies that the ordinal variables are provided in the VAR statement, and the continuous variables are provided in the WITH statement. By default, ORDINAL=WITH.

**RANK**
displays the ordered correlation coefficients for each variable. Correlations are ordered from highest to lowest in absolute value. If you specify the HOEFFDING option, the $D$ statistics are displayed in order from highest to lowest.

**SINGULAR=p**
specifies the criterion for determining the singularity of a variable if you use a PARTIAL statement. A variable is considered singular if its corresponding diagonal element after Cholesky decomposition has a value less than $p$ times the original unpartialled value of that variable. By default, SINGULAR=1E−8. The range of $p$ is between 0 and 1.

**SPEARMAN**
requests a table of Spearman correlation coefficients based on the ranks of the variables. The correlations range from −1 to 1. If you specify a WEIGHT statement, the SPEARMAN option is invalid.

**SSCP**
displays a table of the sums of squares and crossproducts. When you specify the SSCP option, the Pearson correlations are also displayed. If you specify the OUTP= option, the output data set contains a SSCP matrix and the corresponding _TYPE_ variable value is ‘SSCP.’ If you use a PARTIAL statement, the unpartial SSCP matrix is displayed, and the output data set does not contain an SSCP matrix.

**VARDEF=DF | N | WDF | WEIGHT | WGT**
specifies the variance divisor in the calculation of variances and covariances. The default is VARDEF=DF.

Table 2.2 displays available values and associated divisors for the VARDEF= option, where $n$ is the number of nonmissing observations, $k$ is the number of variables specified in the PARTIAL statement, and $w_j$ is the weight associated with the $j$th nonmissing observation.
### Table 2.2 Possible Values for the VARDEF= Option

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
<th>Divisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>Degrees of freedom</td>
<td>$n - k - 1$</td>
</tr>
<tr>
<td>N</td>
<td>Number of observations</td>
<td>$n$</td>
</tr>
<tr>
<td>WDF</td>
<td>Sum of weights minus one</td>
<td>$\sum_{j=1}^{n} w_j - k - 1$</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>Sum of weights</td>
<td>$\sum_{j=1}^{n} w_j$</td>
</tr>
</tbody>
</table>

---

**BY Statement**

```sas
BY variables ;
```

You can specify a BY statement with PROC CORR to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the CORR procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

---

**FREQ Statement**

```sas
FREQ variable ;
```

The FREQ statement lists 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 the FREQ variable. If $n$ is not an integer, SAS truncates it. If $n$ is less than 1 or is missing, the observation is excluded from the analysis. The sum of the frequency variable represents the total number of observations.

The effects of the FREQ and WEIGHT statements are similar except when calculating degrees of freedom.
Chapter 2: The CORR Procedure

**ID Statement**

```plaintext
ID variables ;
```

The ID statement specifies one or more additional tip variables to identify observations in scatter plots and scatter plot matrix. For each plot, the tip variables include the X-axis variable, the Y-axis variable, and the variable for observation numbers.

**PARTIAL Statement**

```plaintext
PARTIAL variables ;
```

The PARTIAL statement lists variables to use in the calculation of partial correlation statistics. Only the Pearson partial correlation, Spearman partial rank-order correlation, and Kendall’s partial tau-b can be computed. When you use the PARTIAL statement, observations with missing values are excluded.

With a PARTIAL statement, PROC CORR also displays the partial variance and standard deviation for each analysis variable if the PEARSON option is specified.

**VAR Statement**

```plaintext
VAR variables ;
```

The VAR statement lists variables for which to compute correlation coefficients. If the VAR statement is not specified, PROC CORR computes correlations for all numeric variables not listed in other statements.

**WEIGHT Statement**

```plaintext
WEIGHT variable ;
```

The WEIGHT statement lists weights to use in the calculation of Pearson weighted product-moment correlation. The HOEFFDING, KENDALL, and SPEARMAN options are not valid with the WEIGHT statement.

The observations with missing weights are excluded from the analysis. By default, for observations with nonpositive weights, weights are set to zero and the observations are included in the analysis. You can use the EXCLNPWGT option to exclude observations with negative or zero weights from the analysis.
WITH Statement

WITH variables;

The WITH statement lists variables with which correlations of the VAR statement variables are to be computed. The WITH statement requests correlations of the form \( r(X_i, Y_j) \), where \( X_1, \ldots, X_m \) are analysis variables specified in the VAR statement, and \( Y_1, \ldots, Y_n \) are variables specified in the WITH statement. The correlation matrix has a rectangular structure of the form

\[
\begin{bmatrix}
  r(Y_1, X_1) & \cdots & r(Y_1, X_m) \\
  \vdots & \ddots & \vdots \\
  r(Y_n, X_1) & \cdots & r(Y_n, X_m)
\end{bmatrix}
\]

For example, the statements

```plaintext
proc corr;
  var x1 x2;
  with y1 y2 y3;
run;
```

produce correlations for the following combinations:

\[
\begin{bmatrix}
  r(Y_1, X_1) & r(Y_1, X_2) \\
  r(Y_2, X_1) & r(Y_2, X_2) \\
  r(Y_3, X_1) & r(Y_3, X_2)
\end{bmatrix}
\]

Details: CORR Procedure

Pearson Product-Moment Correlation

The Pearson product-moment correlation is a parametric measure of association for two variables. It measures both the strength and the direction of a linear relationship. If one variable \( X \) is an exact linear function of another variable \( Y \), a positive relationship exists if the correlation is 1 and a negative relationship exists if the correlation is \(-1\). If there is no linear predictability between the two variables, the correlation is 0. If the two variables are normal with a correlation 0, the two variables are independent. However, correlation does not imply causality because, in some cases, an underlying causal relationship might not exist.
The scatter plot matrix in Figure 2.4 displays the relationship between two numeric random variables in various situations.

**Figure 2.4 Correlations between Two Variables**

The scatter plot matrix shows a positive correlation between variables $Y_1$ and $X_1$, a negative correlation between $Y_1$ and $X_2$, and no clear correlation between $Y_2$ and $X_1$. The plot also shows no clear linear correlation between $Y_2$ and $X_2$, even though $Y_2$ is dependent on $X_2$.

The formula for the population Pearson product-moment correlation, denoted $\rho_{xy}$, is

$$
\rho_{xy} = \frac{\text{Cov}(x, y)}{\sqrt{\text{V}(x)\text{V}(y)}} = \frac{\text{E}(x - \text{E}(x))(y - \text{E}(y))}{\sqrt{\text{E}(x - \text{E}(x))^2 \text{E}(y - \text{E}(y))^2}}
$$
Spearman Rank-Order Correlation

The sample correlation, such as a Pearson product-moment correlation or weighted product-moment correlation, estimates the population correlation. The formula for the sample Pearson product-moment correlation is

\[ r_{xy} = \frac{\sum_i ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \]

where \( \bar{x} \) is the sample mean of \( x \) and \( \bar{y} \) is the sample mean of \( y \). The formula for a weighted Pearson product-moment correlation is

\[ r_{xy} = \frac{\sum_i w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum_i w_i (x_i - \bar{x}_w)^2 \sum_i w_i (y_i - \bar{y}_w)^2}} \]

where \( w_i \) is the weight, \( \bar{x}_w \) is the weighted mean of \( x \), and \( \bar{y}_w \) is the weighted mean of \( y \).

**Probability Values**

Probability values for the Pearson correlation are computed by treating

\[ t = (n - 2)^{1/2} \left( \frac{r^2}{1 - r^2} \right)^{1/2} \]

as coming from a \( t \) distribution with \( (n - 2) \) degrees of freedom, where \( r \) is the sample correlation.

Spearman Rank-Order Correlation

Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values. The formula is

\[ \theta = \frac{\sum_i ((R_i - \bar{R})(S_i - \bar{S}))}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (S_i - \bar{S})^2}} \]

where \( R_i \) is the rank of \( x_i \), \( S_i \) is the rank of \( y_i \), \( \bar{R} \) is the mean of the \( R_i \) values, and \( \bar{S} \) is the mean of the \( S_i \) values.

PROC CORR computes the Spearman correlation by ranking the data and using the ranks in the Pearson product-moment correlation formula. In case of ties, the averaged ranks are used.

**Probability Values**

Probability values for the Spearman correlation are computed by treating

\[ t = (n - 2)^{1/2} \left( \frac{r^2}{1 - r^2} \right)^{1/2} \]

as coming from a \( t \) distribution with \( (n - 2) \) degrees of freedom, where \( r \) is the sample Spearman correlation.
**Kendall’s Tau-b Correlation Coefficient**

Kendall’s tau-b is a nonparametric measure of association based on the number of concordances and discordances in paired observations. Concordance occurs when paired observations vary together, and discordance occurs when paired observations vary differently. The formula for Kendall’s tau-b is

\[
\tau = \frac{\sum_{i<j} (\text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j))}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}
\]

where \(T_0 = n(n - 1)/2\), \(T_1 = \sum_k t_k(t_k - 1)/2\), and \(T_2 = \sum_l u_l(u_l - 1)/2\). The \(t_k\) is the number of tied \(x\) values in the \(k\)th group of tied \(x\) values, \(u_l\) is the number of tied \(y\) values in the \(l\)th group of tied \(y\) values, \(n\) is the number of observations, and \(\text{sgn}(z)\) is defined as

\[
\text{sgn}(z) = \begin{cases} 
1 & \text{if } z > 0 \\
0 & \text{if } z = 0 \\
-1 & \text{if } z < 0
\end{cases}
\]

PROC CORR computes Kendall’s tau-b by ranking the data and using a method similar to Knight (1966). The data are double sorted by ranking observations according to values of the first variable and reranking the observations according to values of the second variable. PROC CORR computes Kendall’s tau-b from the number of interchanges of the first variable and corrects for tied pairs (pairs of observations with equal values of \(X\) or equal values of \(Y\)).

**Probability Values**

Probability values for Kendall’s tau-b are computed by treating

\[
\frac{s}{\sqrt{V(s)}}
\]

as coming from a standard normal distribution where

\[
s = \sum_{i<j} (\text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j))
\]

and \(V(s)\), the variance of \(s\), is computed as

\[
V(s) = \frac{v_0 - v_t - v_u}{18} + \frac{v_1}{2n(n - 1)} + \frac{v_2}{9n(n - 1)(n - 2)}
\]

where

\[
v_0 = n(n - 1)(2n + 5)
\]
\[
v_t = \sum_k t_k(t_k - 1)(2t_k + 5)
\]
\[
v_u = \sum_l u_l(u_l - 1)(2u_l + 5)
\]
\[
v_1 = \left(\sum_k t_k(t_k - 1)\right)\left(\sum u_l(u_l - 1)\right)
\]
\[
v_2 = \left(\sum_l u_l(u_l - 1)\right)\left(\sum t_i(t_i - 2)\right)\left(\sum u_l(u_l - 1)(u_l - 2)\right)
\]

The sums are over tied groups of values where \(t_i\) is the number of tied \(x\) values and \(u_l\) is the number of tied \(y\) values (Noether 1967). The sampling distribution of Kendall’s partial tau-b is unknown; therefore, the probability values are not available.
Hoeffding Dependence Coefficient

Hoeffding’s measure of dependence, $D$, is a nonparametric measure of association that detects more general departures from independence. The statistic approximates a weighted sum over observations of chi-square statistics for two-by-two classification tables (Hoeffding 1948). Each set of $(x, y)$ values are cut points for the classification. The formula for Hoeffding’s $D$ is

$$D = 30 \frac{(n - 2)(n - 3)D_1 + D_2 - 2(n - 2)D_3}{n(n - 1)(n - 2)(n - 3)(n - 4)}$$

where $D_1 = \sum_i (Q_i - 1)(Q_i - 2)$, $D_2 = \sum_i (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$, and $D_3 = \sum_i (R_i - 2)(S_i - 2)(Q_i - 1)$. $R_i$ is the rank of $x_i$, $S_i$ is the rank of $y_i$, and $Q_i$ (also called the bivariate rank) is 1 plus the number of points with both $x$ and $y$ values less than the $i$th point.

A point that is tied on only the $x$ value or $y$ value contributes $1/2$ to $Q_i$ if the other value is less than the corresponding value for the $i$th point.

A point that is tied on both $x$ and $y$ contributes $1/4$ to $Q_i$. PROC CORR obtains the $Q_i$ values by first ranking the data. The data are then double sorted by ranking observations according to values of the first variable and reranking the observations according to values of the second variable. Hoeffding’s $D$ statistic is computed using the number of interchanges of the first variable. When no ties occur among data set observations, the $D$ statistic values are between $-0.5$ and $1$, with $1$ indicating complete dependence. However, when ties occur, the $D$ statistic might result in a smaller value. That is, for a pair of variables with identical values, the Hoeffding’s $D$ statistic might be less than $1$. With a large number of ties in a small data set, the $D$ statistic might be less than $-0.5$. For more information about Hoeffding’s $D$, see Hollander and Wolfe (1999).

Probability Values

The probability values for Hoeffding’s $D$ statistic are computed using the asymptotic distribution computed by Blum, Kiefer, and Rosenblatt (1961). The formula is

$$\frac{(n - 1)\pi^4}{60} - D + \frac{\pi^4}{72}$$

which comes from the asymptotic distribution. If the sample size is less than 10, refer to the tables for the distribution of $D$ in Hollander and Wolfe (1999).

Partial Correlation

A partial correlation measures the strength of a relationship between two variables, while controlling the effect of other variables. The Pearson partial correlation between two variables, after controlling for variables in the PARTIAL statement, is equivalent to the Pearson correlation between the residuals of the two variables after regression on the controlling variables.

Let $y = (y_1, y_2, \ldots, y_v)$ be the set of variables to correlate and $z = (z_1, z_2, \ldots, z_p)$ be the set of controlling variables. The population Pearson partial correlation between the $i$th and the $j$th variables of $y$ given $z$ is the correlation between errors $(y_i - E(y_i))$ and $(y_j - E(y_j))$, where

$$E(y_i) = \alpha_i + z\beta_i \quad \text{and} \quad E(y_j) = \alpha_j + z\beta_j$$

are the regression models for variables $y_i$ and $y_j$ given the set of controlling variables $z$, respectively.
For a given sample of observations, a sample Pearson partial correlation between \( y_i \) and \( y_j \) given \( z \) is derived from the residuals \( y_i - \hat{y}_i \) and \( y_j - \hat{y}_j \), where

\[
\hat{y}_i = \hat{\alpha}_i + z\hat{\beta}_i \quad \text{and} \quad \hat{y}_j = \hat{\alpha}_j + z\hat{\beta}_j
\]

are fitted values from regression models for variables \( y_i \) and \( y_j \) given \( z \).

The partial corrected sums of squares and crossproducts (CSSCP) of \( y \) given \( z \) are the corrected sums of squares and crossproducts of the residuals \( y - \hat{y} \). Using these partial corrected sums of squares and crossproducts, you can calculate the partial covariances and partial correlations.

PROC CORR derives the partial corrected sums of squares and crossproducts matrix by applying the Cholesky decomposition algorithm to the CSSCP matrix. For Pearson partial correlations, let \( S \) be the partitioned CSSCP matrix between two sets of variables, \( z \) and \( y \):

\[
S = \begin{bmatrix}
S_{zz} & S_{zy} \\
S_{zy}' & S_{yy}
\end{bmatrix}
\]

PROC CORR calculates \( S_{yy,z} \), the partial CSSCP matrix of \( y \) after controlling for \( z \), by applying the Cholesky decomposition algorithm sequentially on the rows associated with \( z \), the variables being partialled out.

After applying the Cholesky decomposition algorithm to each row associated with variables \( z \), PROC CORR checks all higher-numbered diagonal elements associated with \( z \) for singularity. A variable is considered singular if the value of the corresponding diagonal element is less than \( \varepsilon \) times the original unpartialled corrected sum of squares of that variable. You can specify the singularity criterion \( \varepsilon \) by using the SINGULAR= option. For Pearson partial correlations, a controlling variable \( z \) is considered singular if the \( R^2 \) for predicting this variable from the variables that are already partialled out exceeds \( 1 - \varepsilon \). When this happens, PROC CORR excludes the variable from the analysis. Similarly, a variable is considered singular if the \( R^2 \) for predicting this variable from the controlling variables exceeds \( 1 - \varepsilon \). When this happens, its associated diagonal element and all higher-numbered elements in this row or column are set to zero.

After the Cholesky decomposition algorithm is applied to all rows associated with \( z \), the resulting matrix has the form

\[
T = \begin{bmatrix}
T_{zz} & T_{zy} \\
0 & S_{yy,z}
\end{bmatrix}
\]

where \( T_{zz} \) is an upper triangular matrix with \( T_{zz}'T_{zz} = S_{zz}' \), \( T_{zz}'T_{zy} = S_{zy}' \), and \( S_{yy,z} = S_{yy} - T_{zy}'T_{zy} \).

If \( S_{zz} \) is positive definite, then \( T_{zy} = T_{zz}'^{-1}S_{zy}' \) and the partial CSSCP matrix \( S_{yy,z} \) is identical to the matrix derived from the formula

\[
S_{yy,z} = S_{yy} - S_{zy}'S_{zz}'^{-1}S_{zy}
\]

The partial variance-covariance matrix is calculated with the variance divisor (VARDEF= option). PROC CORR then uses the standard Pearson correlation formula on the partial variance-covariance matrix to calculate the Pearson partial correlation matrix.

When a correlation matrix is positive definite, the resulting partial correlation between variables \( x \) and \( y \) after adjusting for a single variable \( z \) is identical to that obtained from the first-order partial correlation formula

\[
r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}
\]

where \( r_{xy} \), \( r_{xz} \), and \( r_{yz} \) are the appropriate correlations.
The formula for higher-order partial correlations is a straightforward extension of the preceding first-order formula. For example, when the correlation matrix is positive definite, the partial correlation between \( x \) and \( y \) controlling for both \( z_1 \) and \( z_2 \) is identical to the second-order partial correlation formula

\[
 r_{xy,z_1z_2} = \frac{r_{xy,z_1} - r_{xz_2,z_1}r_{yz_2,z_1}}{\sqrt{(1-r_{xz_2,z_1}^2)(1-r_{yz_2,z_1}^2)}}
\]

where \( r_{xy,z_1} \), \( r_{xz_2,z_1} \), and \( r_{yz_2,z_1} \) are first-order partial correlations among variables \( x, y \), and \( z_2 \) given \( z_1 \).

To derive the corresponding Spearman partial rank-order correlations and Kendall partial tau-b correlations, PROC CORR applies the Cholesky decomposition algorithm to the Spearman rank-order correlation matrix and Kendall’s tau-b correlation matrix and uses the correlation formula. That is, the Spearman partial correlation is equivalent to the Pearson correlation between the residuals of the linear regression of the ranks of the two variables on the ranks of the partialled variables. Thus, if a PARTIAL statement is specified with the CORR=SPEARMAN option, the residuals of the ranks of the two variables are displayed in the plot. The partial tau-b correlations range from –1 to 1. However, the sampling distribution of this partial tau-b is unknown; therefore, the probability values are not available.

**Probability Values**

Probability values for the Pearson and Spearman partial correlations are computed by treating

\[
(n - k - 2)^{1/2}r \\
(1-r^2)^{1/2}
\]

as coming from a \( t \) distribution with \( (n - k - 2) \) degrees of freedom, where \( r \) is the partial correlation and \( k \) is the number of variables being partialled out.

---

**Fisher’s \( z \) Transformation**

For a sample correlation \( r \) that uses a sample from a bivariate normal distribution with correlation \( \rho = 0 \), the statistic

\[
t_r = (n - 2)^{1/2} \left( \frac{r^2}{1-r^2} \right)^{1/2}
\]

has a Student’s \( t \) distribution with \( (n - 2) \) degrees of freedom.

With the monotone transformation of the correlation \( r \) (Fisher 1921)

\[
z_r = \tanh^{-1}(r) = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right)
\]

the statistic \( z \) has an approximate normal distribution with mean and variance

\[
E(z_r) = \zeta + \frac{\rho}{2(n-1)}
\]

\[
V(z_r) = \frac{1}{n-3}
\]

where \( \zeta = \tanh^{-1}(\rho) \).
For the transformed \( z_r \), the approximate variance \( V(z_r) = 1/(n - 3) \) is independent of the correlation \( \rho \). Furthermore, even the distribution of \( z_r \) is not strictly normal, it tends to normality rapidly as the sample size increases for any values of \( \rho \) (Fisher 1973, pp. 200–201).

For the null hypothesis \( H_0: \rho = \rho_0 \), the \( p \)-values are computed by treating

\[
z_r - \zeta_0 = \frac{\rho_0}{2(n - 1)}
\]

as a normal random variable with mean zero and variance \( 1/(n - 3) \), where \( \zeta_0 = \tanh^{-1}(\rho_0) \) (Fisher 1973, p. 207; Anderson 1984, p. 123).

Note that the bias adjustment, \( \rho_0/(2(n - 1)) \), is always used when computing \( p \)-values under the null hypothesis \( H_0: \rho = \rho_0 \) in the CORR procedure.

The ALPHA= option in the FISHER option specifies the value \( \alpha \) for the confidence level \( 1 - \alpha \), the RH0= option specifies the value \( \rho_0 \) in the hypothesis \( H_0: \rho = \rho_0 \), and the BIASADJ= option specifies whether the bias adjustment is to be used for the confidence limits.

The TYPE= option specifies the type of confidence limits. The TYPE=TWOSIDED option requests two-sided confidence limits and a \( p \)-value under the hypothesis \( H_0: \rho = \rho_0 \). For a one-sided confidence limit, the TYPE=LOWER option requests a lower confidence limit and a \( p \)-value under the hypothesis \( H_0: \rho <= \rho_0 \), and the TYPE=UPPER option requests an upper confidence limit and a \( p \)-value under the hypothesis \( H_0: \rho >= \rho_0 \).

### Confidence Limits for the Correlation

The confidence limits for the correlation \( \rho \) are derived through the confidence limits for the parameter \( \zeta \), with or without the bias adjustment.

Without a bias adjustment, confidence limits for \( \zeta \) are computed by treating

\[
z_r - \zeta
\]

as having a normal distribution with mean zero and variance \( 1/(n - 3) \).

That is, the two-sided confidence limits for \( \zeta \) are computed as

\[
\zeta_l = z_r - z_{(1-\alpha/2)} \sqrt{\frac{1}{n - 3}}
\]

\[
\zeta_u = z_r + z_{(1-\alpha/2)} \sqrt{\frac{1}{n - 3}}
\]

where \( z_{(1-\alpha/2)} \) is the \( 100(1 - \alpha/2) \) percentage point of the standard normal distribution.

With a bias adjustment, confidence limits for \( \zeta \) are computed by treating

\[
z_r - \zeta - \text{bias}(r)
\]

as having a normal distribution with mean zero and variance \( 1/(n - 3) \), where the bias adjustment function (Keeping 1962, p. 308) is

\[
\text{bias}(r_r) = \frac{r}{2(n - 1)}
\]
That is, the two-sided confidence limits for $\xi$ are computed as

$$
\xi_l = z_r - \text{bias}(r) - z_{1-\alpha/2} \sqrt{\frac{1}{n - 3}}
$$

$$
\xi_u = z_r - \text{bias}(r) + z_{1-\alpha/2} \sqrt{\frac{1}{n - 3}}
$$

These computed confidence limits of $\xi_l$ and $\xi_u$ are then transformed back to derive the confidence limits for the correlation $\rho$:

$$
r_l = \tanh(\xi_l) = \frac{\exp(2\xi_l) - 1}{\exp(2\xi_l) + 1}
$$

$$
r_u = \tanh(\xi_u) = \frac{\exp(2\xi_u) - 1}{\exp(2\xi_u) + 1}
$$

Note that with a bias adjustment, the CORR procedure also displays the following correlation estimate:

$$
r_{adj} = \tanh(z_r - \text{bias}(r))
$$

### Applications of Fisher’s $z$ Transformation

Fisher (1973, p. 199) describes the following practical applications of the $z$ transformation:

- testing whether a population correlation is equal to a given value
- testing for equality of two population correlations
- combining correlation estimates from different samples

To test if a population correlation $\rho_1$ from a sample of $n_1$ observations with sample correlation $r_1$ is equal to a given $\rho_0$, first apply the $z$ transformation to $r_1$ and $\rho_0$: $z_1 = \tanh^{-1}(r_1)$ and $\xi_0 = \tanh^{-1}(\rho_0)$.

The $p$-value is then computed by treating

$$
z_1 - \xi_0 - \frac{\rho_0}{2(n_1 - 1)}
$$

as a normal random variable with mean zero and variance $1/(n_1 - 3)$.

Assume that sample correlations $r_1$ and $r_2$ are computed from two independent samples of $n_1$ and $n_2$ observations, respectively. To test whether the two corresponding population correlations, $\rho_1$ and $\rho_2$, are equal, first apply the $z$ transformation to the two sample correlations: $z_1 = \tanh^{-1}(r_1)$ and $z_2 = \tanh^{-1}(r_2)$.

The $p$-value is derived under the null hypothesis of equal correlation. That is, the difference $z_1 - z_2$ is distributed as a normal random variable with mean zero and variance $1/(n_1 - 3) + 1/(n_2 - 3)$.

Assuming further that the two samples are from populations with identical correlation, a combined correlation estimate can be computed. The weighted average of the corresponding $z$ values is

$$
\bar{z} = \frac{(n_1 - 3)z_1 + (n_2 - 3)z_2}{n_1 + n_2 - 6}
$$

where the weights are inversely proportional to their variances.

Thus, a combined correlation estimate is $\bar{r} = \tanh(\bar{z})$ and $V(\bar{z}) = 1/(n_1 + n_2 - 6)$. See Example 2.4 for further illustrations of these applications.

Note that this approach can be extended to include more than two samples.
Polychoric Correlation

Polychoric correlation measures the correlation between two unobserved, continuous variables that have a bivariate normal distribution. Information about each unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values (Olsson 1979; Drasgow 1986).

The polychoric correlation coefficient is the maximum likelihood estimate of the product-moment correlation between the underlying normal variables. The range of the polychoric correlation is from –1 to 1. Olsson (1979) gives the likelihood equations and the asymptotic standard errors for estimating the polychoric correlation. The underlying continuous variables relate to the observed ordinal variables through thresholds, which define a range of numeric values that correspond to each categorical level. PROC CORR uses Olsson’s maximum likelihood method for simultaneous estimation of the polychoric correlation and the thresholds.

PROC CORR iteratively solves the likelihood equations by using a Newton-Raphson algorithm. The initial estimates of the thresholds are computed from the inverse of the normal distribution function at the cumulative marginal proportions of the table. Iterative computation of the polychoric correlation stops when the convergence measure falls below the convergence criterion or when the maximum number of iterations is reached, whichever occurs first.

Probability Values

The CORR procedure computes two types of testing for the zero polychoric correlation: the Wald test and the likelihood ratio (LR) test.

Given the maximum likelihood estimate of the polychoric correlation \( \hat{\rho} \) and its asymptotic standard error \( \text{StdErr}(\hat{\rho}) \), the Wald chi-square test statistic is computed as

\[
\left( \frac{\hat{\rho}}{\text{StdErr}(\hat{\rho})} \right)^2
\]

The Wald statistic has an asymptotic chi-square distribution with one degree of freedom.

For the LR test, the maximum likelihood function assuming zero polychoric correlation is also needed. The LR test statistic is computed as

\[
-2 \log \left( \frac{L_0}{L_1} \right)
\]

where \( L_1 \) is the likelihood function with the maximum likelihood estimates for all parameters, and \( L_0 \) is the likelihood function with the maximum likelihood estimates for all parameters except the polychoric correlation, which is set to 0. The LR statistic also has an asymptotic chi-square distribution with one degree of freedom.
Polyserial Correlation

Polyserial correlation measures the correlation between two continuous variables with a bivariate normal distribution, where one variable is observed directly, and the other is unobserved. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values (Olsson, Drasgow, and Dorans 1982).

Let $X$ be the observed continuous variable from a normal distribution with mean $\mu$ and variance $\sigma^2$, let $Y$ be the unobserved continuous variable, and let $\rho$ be the Pearson correlation between $X$ and $Y$. Furthermore, assume that an observed ordinal variable $D$ is derived from $Y$ as follows:

$$
D = \begin{cases} 
  d^{(1)} & \text{if } Y < \tau_1 \\
  d^{(k)} & \text{if } \tau_{k-1} \leq Y < \tau_k, \ k = 2, 3, \ldots, K - 1 \\
  d^{(K)} & \text{if } Y \geq \tau_{K-1}
\end{cases}
$$

where $d^{(1)} < d^{(2)} < \ldots < d^{(K)}$ are ordered observed values, and $\tau_1 < \tau_2 < \ldots < \tau_{K-1}$ are ordered unknown threshold values.

The likelihood function for the joint distribution $(X, D)$ from a sample of $N$ observations $(x_j, d_j)$ is

$$
L = \prod_{j=1}^{N} f(x_j, d_j) = \prod_{j=1}^{N} f(x_j) \ P(D = d_j \mid x_j)
$$

where $f(x_j)$ is the normal density function with mean $\mu$ and standard deviation $\sigma$ (Drasgow 1986).

The conditional distribution of $Y$ given $X = x_j$ is normal with mean $\rho z_j$ and variance $1 - \rho^2$, where $z_j = (x_j - \mu)/\sigma$ is a standard normal variate. Without loss of generality, assume the variable $Y$ has a standard normal distribution. Then if $d_j = d^{(k)}$, the $k^{th}$ ordered value in $D$, the resulting conditional density is

$$
P(D = d^{(k)} \mid x_j) = \begin{cases} 
  \Phi \left( \frac{\tau_1 - \rho z_j}{\sqrt{1 - \rho^2}} \right) & \text{if } k = 1 \\
  \Phi \left( \frac{\tau_k - \rho z_j}{\sqrt{1 - \rho^2}} \right) - \Phi \left( \frac{\tau_{k-1} - \rho z_j}{\sqrt{1 - \rho^2}} \right) & \text{if } k = 2, 3, \ldots, K - 1 \\
  1 - \Phi \left( \frac{\tau_{K-1} - \rho z_j}{\sqrt{1 - \rho^2}} \right) & \text{if } k = K
\end{cases}
$$

where $\Phi$ is the cumulative normal distribution function.

Cox (1974) derives the maximum likelihood estimates for all parameters $\mu$, $\sigma$, $\rho$ and $\tau_1, \ldots, \tau_{K-1}$. The maximum likelihood estimates for $\mu$ and $\sigma^2$ can be derived explicitly. The maximum likelihood estimate for $\mu$ is the sample mean and the maximum likelihood estimate for $\sigma^2$ is the sample variance

$$
\sigma^2 = \frac{\sum_{j=1}^{N} (x_j - \bar{x})^2}{N}
$$

The maximum likelihood estimates for the remaining parameters, including the polyserial correlation $\rho$ and thresholds $\tau_1, \ldots, \tau_{K-1}$, can be computed by an iterative process, as described by Cox (1974). The asymptotic standard error of the maximum likelihood estimate of $\rho$ can also be computed after this process.
For a vector of parameters, the information matrix is the negative of the Hessian matrix (the matrix of second partial derivatives of the log likelihood function), and is used in the computation of the maximum likelihood estimates of these parameters. The CORR procedure uses the observed information matrix (the information matrix evaluated at the current parameter estimates) in the computation. After the maximum likelihood estimates are derived, the asymptotic covariance matrix for these parameter estimates is computed as the inverse of the observed information matrix (the information matrix evaluated at the maximum likelihood estimates).

**Probability Values**

The CORR procedure computes two types of testing for the zero polyserial correlation: the Wald test and the likelihood ratio (LR) test.

Given the maximum likelihood estimate of the polyserial correlation $\hat{\rho}$ and its asymptotic standard error $\text{StdErr}(\hat{\rho})$, the Wald chi-square test statistic is computed as

$$\left(\frac{\hat{\rho}}{\text{StdErr}(\hat{\rho})}\right)^2$$

The Wald statistic has an asymptotic chi-square distribution with one degree of freedom.

For the LR test, the maximum likelihood function assuming zero polyserial correlation is also needed. If $\rho = 0$, the likelihood function is reduced to

$$L = \prod_{j=1}^{N} f(x_j, d_j) = \prod_{j=1}^{N} f(x_j) \prod_{j=1}^{N} P(D = d_j)$$

In this case, the maximum likelihood estimates for all parameters can be derived explicitly. The maximum likelihood estimates for $\mu$ is the sample mean and the maximum likelihood estimate for $\sigma^2$ is the sample variance

$$\frac{\sum_{j=1}^{N} (x_j - \bar{x})^2}{N}$$

In addition, the maximum likelihood estimate for the threshold $\tau_k$, $k = 1, \ldots, K - 1$, is

$$\Phi^{-1} \left( \frac{\sum_{g=1}^{k} n_g}{N} \right)$$

where $n_g$ is the number of observations in the $g^{th}$ ordered group of the ordinal variable $D$, and $N = \sum_{g=1}^{K} n_g$ is the total number of observations.

The LR test statistic is computed as

$$-2 \log \left( \frac{L_0}{L_1} \right)$$

where $L_1$ is the likelihood function with the maximum likelihood estimates for all parameters, and $L_0$ is the likelihood function with the maximum likelihood estimates for all parameters except the polyserial correlation, which is set to 0. The LR statistic also has an asymptotic chi-square distribution with one degree of freedom.
Cronbach’s Coefficient Alpha

Analyzing latent constructs such as job satisfaction, motor ability, sensory recognition, or customer satisfaction requires instruments to accurately measure the constructs. Interrelated items can be summed to obtain an overall score for each participant. Cronbach’s coefficient alpha estimates the reliability of this type of scale by determining the internal consistency of the test or the average correlation of items within the test (Cronbach 1951).

When a value is recorded, the observed value contains some degree of measurement error. Two sets of measurements on the same variable for the same individual might not have identical values. However, repeated measurements for a series of individuals will show some consistency. Reliability measures internal consistency from one set of measurements to another. The observed value \( Y \) is divided into two components, a true value \( T \) and a measurement error \( E \). The measurement error is assumed to be independent of the true value; that is,

\[
Y = T + E \quad \text{Cov}(T, E) = 0
\]

The reliability coefficient of a measurement test is defined as the squared correlation between the observed value \( Y \) and the true value \( T \); that is,

\[
r^2(Y, T) = \frac{\text{Cov}(Y, T)^2}{V(Y)V(T)} = \frac{V(T)^2}{V(Y)V(T)} = \frac{V(T)}{V(Y)}
\]

which is the proportion of the observed variance due to true differences among individuals in the sample. If \( Y \) is the sum of several observed variables measuring the same feature, you can estimate \( V(T) \). Cronbach’s coefficient alpha, based on a lower bound for \( V(T) \), is an estimate of the reliability coefficient.

Suppose \( p \) variables are used with \( Y_j = T_j + E_j \) for \( j = 1, 2, \ldots, p \), where \( Y_j \) is the observed value, \( T_j \) is the true value, and \( E_j \) is the measurement error. The measurement errors \( (E_j) \) are independent of the true values \( (T_j) \) and are also independent of each other. Let \( Y_0 = \sum_j Y_j \) be the total observed score and let \( T_0 = \sum_j T_j \) be the total true score. Because

\[
(p - 1) \sum_j V(T_j) \geq \sum_{i \neq j} \text{Cov}(T_i, T_j)
\]

a lower bound for \( V(T_0) \) is given by

\[
\frac{p}{p-1} \sum_{i \neq j} \text{Cov}(T_i, T_j)
\]

With \( \text{Cov}(Y_i, Y_j) = \text{Cov}(T_i, T_j) \) for \( i \neq j \), a lower bound for the reliability coefficient, \( V(T_0)/V(Y_0) \), is then given by the Cronbach’s coefficient alpha:

\[
\alpha = \left( \frac{p}{p-1} \right) \sum_{i \neq j} \frac{\text{Cov}(Y_i, Y_j)}{V(Y_0)} = \left( \frac{p}{p-1} \right) \left( 1 - \frac{\sum_j V(Y_j)}{V(Y_0)} \right)
\]

If the variances of the items vary widely, you can standardize the items to a standard deviation of 1 before computing the coefficient alpha. If the variables are dichotomous \((0,1)\), the coefficient alpha is equivalent to the Kuder-Richardson 20 (KR-20) reliability measure.
When the correlation between each pair of variables is 1, the coefficient alpha has a maximum value of 1. With negative correlations between some variables, the coefficient alpha can have a value less than zero. The larger the overall alpha coefficient, the more likely that items contribute to a reliable scale. Nunnally and Bernstein (1994) suggests 0.70 as an acceptable reliability coefficient; smaller reliability coefficients are seen as inadequate. However, this varies by discipline.

To determine how each item reflects the reliability of the scale, you calculate a coefficient alpha after deleting each variable independently from the scale. Cronbach’s coefficient alpha from all variables except the $k$th variable is given by

$$\alpha_k = \left( \frac{p - 1}{p - 2} \right) \left( 1 - \frac{\sum_{i \neq k} V(Y_i)}{V(\sum_{i \neq k} Y_i)} \right)$$

If the reliability coefficient increases after an item is deleted from the scale, you can assume that the item is not correlated highly with other items in the scale. Conversely, if the reliability coefficient decreases, you can assume that the item is highly correlated with other items in the scale. Refer to Yu (2001) for more information about how to interpret Cronbach’s coefficient alpha.

Listwise deletion of observations with missing values is necessary to correctly calculate Cronbach’s coefficient alpha. PROC CORR does not automatically use listwise deletion if you specify the ALPHA option. Therefore, you should use the NOMISS option if the data set contains missing values. Otherwise, PROC CORR prints a warning message indicating the need to use the NOMISS option with the ALPHA option.

Confidence and Prediction Ellipses

When the relationship between two variables is nonlinear or when outliers are present, the correlation coefficient might incorrectly estimate the strength of the relationship. Plotting the data enables you to verify the linear relationship and to identify the potential outliers.

The partial correlation between two variables, after controlling for variables in the PARTIAL statement, is the correlation between the residuals of the linear regression of the two variables on the partialled variables. Thus, if a PARTIAL statement is also specified, the residuals of the analysis variables are displayed in the scatter plot matrix and scatter plots.

The CORR procedure optionally provides two types of ellipses for each pair of variables in a scatter plot. One is a confidence ellipse for the population mean, and the other is a prediction ellipse for a new observation. Both assume a bivariate normal distribution.

Let $\boldsymbol{Z}$ and $\mathbf{S}$ be the sample mean and sample covariance matrix of a random sample of size $n$ from a bivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. The variable $\boldsymbol{Z} - \boldsymbol{\mu}$ is distributed as a bivariate normal variate with mean zero and covariance $(1/n)\mathbf{\Sigma}$, and it is independent of $\mathbf{S}$. Using Hotelling’s $T^2$ statistic, which is defined as

$$T^2 = n(\boldsymbol{Z} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\boldsymbol{Z} - \boldsymbol{\mu})$$

a $100(1 - \alpha)$% confidence ellipse for $\boldsymbol{\mu}$ is computed from the equation

$$\frac{n}{n - 1} (\bar{\boldsymbol{Z}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\boldsymbol{Z}} - \boldsymbol{\mu}) = \frac{2}{n - 2} F_{2,n-2}(1 - \alpha)$$

where $F_{2,n-2}(1 - \alpha)$ is the $(1 - \alpha)$ critical value of an $F$ distribution with degrees of freedom 2 and $n - 2$. 

A prediction ellipse is a region for predicting a new observation in the population. It also approximates a region that contains a specified percentage of the population.

Denote a new observation as the bivariate random variable $Z_{\text{new}}$. The variable 

$$Z_{\text{new}} - \bar{Z} = (Z_{\text{new}} - \mu) - (\bar{Z} - \mu)$$

is distributed as a bivariate normal variate with mean zero (the zero vector) and covariance $(1 + 1/n) \Sigma$, and it is independent of $S$. A $100(1 - \alpha)\%$ prediction ellipse is then given by the equation

$$\frac{n}{n-1}(\bar{Z} - \mu)'S^{-1}(\bar{Z} - \mu) = \frac{2(n + 1)}{n-2} F_{2,n-2}(1 - \alpha)$$

The family of ellipses generated by different critical values of the $F$ distribution has a common center (the sample mean) and common major and minor axis directions.

The shape of an ellipse depends on the aspect ratio of the plot. The ellipse indicates the correlation between the two variables if the variables are standardized (by dividing the variables by their respective standard deviations). In this situation, the ratio between the major and minor axis lengths is

$$\sqrt{\frac{1 + |r|}{1 - |r|}}$$

In particular, if $r = 0$, the ratio is 1, which corresponds to a circular confidence contour and indicates that the variables are uncorrelated. A larger value of the ratio indicates a larger positive or negative correlation between the variables.

**Missing Values**

PROC CORR excludes observations with missing values in the WEIGHT and FREQ variables. By default, PROC CORR uses **pairwise deletion** when observations contain missing values. PROC CORR includes all nonmissing pairs of values for each pair of variables in the statistical computations. Therefore, the correlation statistics might be based on different numbers of observations.

If you specify the NOMISS option, PROC CORR uses **listwise deletion** when a value of the VAR or WITH statement variable is missing. PROC CORR excludes all observations with missing values from the analysis. Therefore, the number of observations for each pair of variables is identical.

The PARTIAL statement always excludes the observations with missing values by automatically invoking the NOMISS option. With the NOMISS option, the data are processed more efficiently because fewer resources are needed. Also, the resulting correlation matrix is nonnegative definite.

In contrast, if the data set contains missing values for the analysis variables and the NOMISS option is not specified, the resulting correlation matrix might not be nonnegative definite. This leads to several statistical difficulties if you use the correlations as input to regression or other statistical procedures.
In-Database Computation

The CORR procedure can use in-database computation to compute univariate statistics and the SSCP matrix if the DATA= input data set is stored as a table in a database management system (DBMS). When the CORR procedure performs in-database computation for the DATA= data set, the procedure generates an SQL query that computes summary tables of univariate statistics and the SSCP matrix. The query is passed to the DBMS and executed in-database. The results of the query are then passed back to the SAS System and transmitted to PROC CORR. The CORR procedure then uses these summary tables to perform the remaining tasks (such as producing the correlation and covariance matrices) in the usual way (out of the database).

In-database computation can provide the advantages of faster processing and reduced data transfer between the database and SAS software. For information about in-database computation, see the section “In-Database Procedures” in Sas/ACCESS for Relational Databases: Reference. Instead of transferring the entire data set over the network between the database and SAS software, the in-database method transfers only the summary tables. This can substantially reduce processing time when the dimensions of the summary tables (in terms of rows and columns) are much smaller than the dimensions of the entire database table (in terms of individual observations). Additionally, in-database summarization uses efficient parallel processing, which can also provide performance advantages.

By default, PROC CORR uses in-database computation when possible. If in-database computation is used, the EXCLNPWGRT option is activated to exclude observations with nonpositive weights. The ID statement requires row-level access and therefore cannot be used in-database. In addition, the HOEFFDING, KENDALL, SPEARMAN, OUTH=, OUTK=, OUTS=, and PLOTS= options also require row-level access and cannot be used in-database.

In-database computation is controlled by the SQLGENERATION option, which you can specify in either a LIBNAME statement or an OPTIONS statement. See the section “In-Database Procedures” in SAS/ACCESS for Relational Databases: Reference for details about the SQLGENERATION option and other options that affect in-database computation. There are no CORR procedure options that control in-database computation.

The order of observations is not inherently defined for DBMS tables. The following options relate to the order of observations and therefore should not be specified for PROC CORR in-database computation:

- If you specify the FIRSTOBS= or OBS= data set option, PROC CORR does not perform in-database computation.

- If you specify the NOTSORTED option in the BY statement, PROC CORR in-database computation ignores it and uses the default ASCENDING order for BY variables.

**NOTE:** In-database computing in the CORR procedure requires installation of the SAS Analytics Accelerator.

Output Tables

By default, PROC CORR prints a report that includes descriptive statistics and correlation statistics for each variable. The descriptive statistics include the number of observations with nonmissing values, the mean, the standard deviation, the minimum, and the maximum.
If a nonparametric measure of association is requested, the descriptive statistics include the median. Otherwise, the sample sum is included. If a Pearson partial correlation is requested, the descriptive statistics also include the partial variance and partial standard deviation.

If variable labels are available, PROC CORR labels the variables. If you specify the CSSCP, SSCP, or COV option, the appropriate sums of squares and crossproducts and covariance matrix appear at the top of the correlation report. If the data set contains missing values, PROC CORR prints additional statistics for each pair of variables. These statistics, calculated from the observations with nonmissing row and column variable values, might include the following:

- SSCP(‘W’,’V’), uncorrected sums of squares and crossproducts
- USS(‘W’), uncorrected sums of squares for the row variable
- USS(‘V’), uncorrected sums of squares for the column variable
- CSSCP(‘W’,’V’), corrected sums of squares and crossproducts
- CSS(‘W’), corrected sums of squares for the row variable
- CSS(‘V’), corrected sums of squares for the column variable
- COV(‘W’,’V’), covariance
- VAR(‘W’), variance for the row variable
- VAR(‘V’), variance for the column variable
- DF(‘W’,’V’), divisor for calculating covariance and variances

For each pair of variables, PROC CORR prints the correlation coefficients, the number of observations used to calculate the coefficient, and the $p$-value.

If you specify the ALPHA option, PROC CORR prints Cronbach’s coefficient alpha, the correlation between the variable and the total of the remaining variables, and Cronbach’s coefficient alpha by using the remaining variables for the raw variables and the standardized variables.

Output Data Sets

If you specify the OUTP=, OUTS=, OUTK=, or OUTH= option, PROC CORR creates an output data set that contains statistics for Pearson correlation, Spearman correlation, Kendall’s tau-b, or Hoeffding’s $D$, respectively. By default, the output data set is a special data set type (TYPE=CORR) that many SAS/STAT procedures recognize, including PROC REG and PROC FACTOR. When you specify the NOCORR option and the COV, CSSCP, or SSCP option, use the TYPE= data set option to change the data set type to COV, CSSCP, or SSCP.

The output data set includes the following variables:

- BY variables, which identify the BY group when using a BY statement
- _TYPE_ variable, which identifies the type of observation
• _NAME_ variable, which identifies the variable that corresponds to a given row of the correlation matrix

• INTERCEPT variable, which identifies variable sums when specifying the SSCP option

• VAR variables, which identify the variables listed in the VAR statement

You can use a combination of the _TYPE_ and _NAME_ variables to identify the contents of an observation. The _NAME_ variable indicates which row of the correlation matrix the observation corresponds to. The values of the _TYPE_ variable are as follows:

• SSCP, uncorrected sums of squares and crossproducts
• CSSCP, corrected sums of squares and crossproducts
• COV, covariances
• MEAN, mean of each variable
• STD, standard deviation of each variable
• N, number of nonmissing observations for each variable
• SUMWGT, sum of the weights for each variable when using a WEIGHT statement
• CORR, correlation statistics for each variable

If you specify the SSCP option, the OUTP= data set includes an additional observation that contains intercept values. If you specify the ALPHA option, the OUTP= data set also includes observations with the following _TYPE_ values:

• RAWALPHA, Cronbach’s coefficient alpha for raw variables
• STDALPHA, Cronbach’s coefficient alpha for standardized variables
• RAWALDEL, Cronbach’s coefficient alpha for raw variables after deleting one variable
• STDALDEL, Cronbach’s coefficient alpha for standardized variables after deleting one variable
• RAWCTDEL, the correlation between a raw variable and the total of the remaining raw variables
• STDCTDEL, the correlation between a standardized variable and the total of the remaining standardized variables

If you use a PARTIAL statement, the statistics are calculated after the variables are partialled. If PROC CORR computes Pearson correlation statistics, MEAN equals zero and STD equals the partial standard deviation associated with the partial variance for the OUTP=, OUTK=, and OUTS= data sets. Otherwise, PROC CORR assigns missing values to MEAN and STD.
PROC CORR assigns a name to each table it creates. You must use these names to reference tables when using the Output Delivery System (ODS). These names are listed in Table 2.3 and Table 2.4. For more information about ODS, see Chapter 20, “Using the Output Delivery System” (SAS/STAT User’s Guide).

### Table 2.3 ODS Tables Produced by PROC CORR

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cov</td>
<td>Covariances</td>
<td>COV</td>
</tr>
<tr>
<td>CronbachAlpha</td>
<td>Coefficient alpha</td>
<td>ALPHA</td>
</tr>
<tr>
<td>CronbachAlphaDel</td>
<td>Coefficient alpha with deleted variable</td>
<td>ALPHA</td>
</tr>
<tr>
<td>Csscp</td>
<td>Corrected sums of squares and crossproducts</td>
<td>CSSCP</td>
</tr>
<tr>
<td>FisherPearsonCorr</td>
<td>Pearson correlation statistics using FISHER</td>
<td>FISHER</td>
</tr>
<tr>
<td>FisherSpearmanCorr</td>
<td>Spearman correlation statistics using Fisher’s z transformation</td>
<td>FISHER SPEARMAN</td>
</tr>
<tr>
<td>HoeffdingCorr</td>
<td>Hoeffding’s D statistics</td>
<td>HOEFFDING</td>
</tr>
<tr>
<td>KendallCorr</td>
<td>Kendall’s tau-b coefficients</td>
<td>KENDALL</td>
</tr>
<tr>
<td>PearsonCorr</td>
<td>Pearson correlations</td>
<td>PEARSON</td>
</tr>
<tr>
<td>PolychoricCorr</td>
<td>Polychoric correlations</td>
<td>POLYCHORIC</td>
</tr>
<tr>
<td>PolyserialCorr</td>
<td>Polyserial correlations</td>
<td>POLYSERIAL</td>
</tr>
<tr>
<td>SimpleStats</td>
<td>Simple descriptive statistics</td>
<td></td>
</tr>
<tr>
<td>SpearmanCorr</td>
<td>Spearman correlations</td>
<td>SPEARMAN</td>
</tr>
<tr>
<td>Sscp</td>
<td>Sums of squares and crossproducts</td>
<td>SSCP</td>
</tr>
<tr>
<td>VarInformation</td>
<td>Variable information</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2.4 ODS Tables Produced with the PARTIAL Statement

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>FisherPearsonPartialCorr</td>
<td>Pearson partial correlation statistics using Fisher’s z transformation</td>
<td>FISHER</td>
</tr>
<tr>
<td>FisherSpearmanPartialCorr</td>
<td>Spearman partial correlation statistics using Fisher’s z transformation</td>
<td>FISHER SPEARMAN</td>
</tr>
<tr>
<td>PartialCsscp</td>
<td>Partial corrected sums of squares and crossproducts</td>
<td>CSSCP</td>
</tr>
<tr>
<td>PartialCov</td>
<td>Partial covariances</td>
<td>COV</td>
</tr>
<tr>
<td>PartialKendallCorr</td>
<td>Partial Kendall tau-b coefficients</td>
<td>KENDALL</td>
</tr>
<tr>
<td>PartialPearsonCorr</td>
<td>Partial Pearson correlations</td>
<td></td>
</tr>
<tr>
<td>PartialSpearmanCorr</td>
<td>Partial Spearman correlations</td>
<td>SPEARMAN</td>
</tr>
</tbody>
</table>


**ODS Graphics**


Before you create graphs, ODS Graphics must be enabled (for example, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” in that chapter.

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “A Primer on ODS Statistical Graphics” in that chapter.

PROC CORR assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. To request these graphs, ODS Graphics must be enabled and you must specify the options indicated in Table 2.5.

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScatterPlot</td>
<td>Scatter plot</td>
<td>PLOTS=SCATTER</td>
</tr>
<tr>
<td>MatrixPlot</td>
<td>Scatter plot matrix</td>
<td>PLOTS=MATRIX</td>
</tr>
</tbody>
</table>

**Examples: CORR Procedure**

**Example 2.1: Computing Four Measures of Association**

This example produces a correlation analysis with descriptive statistics and four measures of association: the Pearson product-moment correlation, the Spearman rank-order correlation, Kendall’s tau-b coefficients, and Hoeffding’s measure of dependence, $D$.

The Fitness data set created in the section “Getting Started: CORR Procedure” on page 5 contains measurements from a study of physical fitness of 31 participants. The following statements request all four measures of association for the variables Weight, Oxygen, and Runtime:

```sas
ods graphics on;
title 'Measures of Association for a Physical Fitness Study';
proc corr data=Fitness pearson spearman kendall hoeffding
   plots=matrix(histogram);
   var Weight Oxygen RunTime;
run;
ods graphics off;
```

Note that Pearson correlations are computed by default only if all three nonparametric correlations (SPEARMAN, KENDALL, and HOEFFDING) are not specified. Otherwise, you need to specify the PEARSON option explicitly to compute Pearson correlations.
The “Simple Statistics” table in Output 2.1.1 displays univariate descriptive statistics for analysis variables. By default, observations with nonmissing values for each variable are used to derive the univariate statistics for that variable. When nonparametric measures of association are specified, the procedure displays the median instead of the sum as an additional descriptive measure.

**Output 2.1.1** Simple Statistics

<table>
<thead>
<tr>
<th>Measures of Association for a Physical Fitness Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>The CORR Procedure</td>
</tr>
<tr>
<td>3 Variables: Weight Oxygen RunTime</td>
</tr>
</tbody>
</table>

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>31</td>
<td>77.44</td>
<td>8.32</td>
<td>77.45</td>
<td>59.08</td>
<td>91.63</td>
</tr>
<tr>
<td>Oxygen</td>
<td>29</td>
<td>47.23</td>
<td>5.48</td>
<td>46.67</td>
<td>37.39</td>
<td>60.06</td>
</tr>
<tr>
<td>RunTime</td>
<td>29</td>
<td>10.67</td>
<td>1.39</td>
<td>10.50</td>
<td>8.17</td>
<td>14.03</td>
</tr>
</tbody>
</table>

The “Pearson Correlation Coefficients” table in Output 2.1.2 displays Pearson correlation statistics for pairs of analysis variables. The Pearson correlation is a parametric measure of association for two continuous random variables. When there are missing data, the number of observations used to calculate the correlation can vary.

**Output 2.1.2** Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th>Pearson Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob &gt;</td>
</tr>
<tr>
<td>Number of Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Oxygen</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1.00000</td>
<td>-0.15358</td>
<td>0.20072</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Oxygen</td>
<td>-0.15358</td>
<td>1.00000</td>
<td>-0.86843</td>
</tr>
<tr>
<td></td>
<td>0.4264</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RunTime</td>
<td>0.20072</td>
<td>-0.86843</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.2965</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>

The table shows that the Pearson correlation between Runtime and Oxygen is $-0.86843$, which is significant with a $p$-value less than 0.0001. This indicates a strong negative linear relationship between these two variables. As Runtime increases, Oxygen decreases linearly.
The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values. The “Spearman Correlation Coefficients” table in Output 2.1.3 displays results similar to those of the “Pearson Correlation Coefficients” table in Output 2.1.2.

**Output 2.1.3** Spearman Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Oxygen</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1.00000</td>
<td>-0.06824</td>
<td>0.13749</td>
</tr>
<tr>
<td></td>
<td>0.7250</td>
<td>0.4769</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Oxygen</td>
<td>-0.06824</td>
<td>1.00000</td>
<td>-0.80131</td>
</tr>
<tr>
<td></td>
<td>0.7250</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>RunTime</td>
<td>0.13749</td>
<td>-0.80131</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.4769</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>

Kendall’s tau-b is a nonparametric measure of association based on the number of concordances and discordances in paired observations. The “Kendall Tau b Correlation Coefficients” table in Output 2.1.4 displays results similar to those of the “Pearson Correlation Coefficients” table in Output 2.1.2.

**Output 2.1.4** Kendall’s Tau-b Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Oxygen</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1.00000</td>
<td>-0.00988</td>
<td>0.06675</td>
</tr>
<tr>
<td></td>
<td>0.9402</td>
<td>0.6123</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Oxygen</td>
<td>-0.00988</td>
<td>1.00000</td>
<td>-0.62434</td>
</tr>
<tr>
<td></td>
<td>0.9402</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>RunTime</td>
<td>0.06675</td>
<td>-0.62434</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.6123</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>
Hoeffding’s measure of dependence, $D$, is a nonparametric measure of association that detects more general departures from independence. Without ties in the variables, the values of the $D$ statistic can vary between $-0.5$ and $1$, with $1$ indicating complete dependence. Otherwise, the $D$ statistic can result in a smaller value. The “Hoeffding Dependence Coefficients” table in Output 2.1.5 displays Hoeffding dependence coefficients. Since ties occur in the variable Weight, the $D$ statistic for the Weight variable is less than $1$.

**Output 2.1.5** Hoeffding’s Dependence Coefficients

<table>
<thead>
<tr>
<th>Hoeffding Dependence Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob &gt; D under H0: D=0</td>
</tr>
<tr>
<td>Number of Observations</td>
</tr>
<tr>
<td>Weight</td>
</tr>
<tr>
<td>Weight</td>
</tr>
<tr>
<td>Oxygen</td>
</tr>
<tr>
<td>RunTime</td>
</tr>
</tbody>
</table>
When you use the PLOTS=MATRIX(HISTOGRAM) option, the CORR procedure displays a symmetric matrix plot for the analysis variables listed in the VAR statement (Output 2.1.6).

**Output 2.1.6** Symmetric Scatter Plot Matrix

The strong negative linear relationship between Oxygen and Runtime is evident in Output 2.1.6.

Note that this graphical display is requested by enabling ODS Graphics and by specifying the PLOTS= option. For more information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS” (SAS/STAT User’s Guide).
Example 2.2: Computing Correlations between Two Sets of Variables

The following statements create the data set *Setosa*, which contains measurements for four iris parts from Fisher’s iris data (1936): sepal length, sepal width, petal length, and petal width. The data set has been altered to contain some missing values.

```
*------------------- Data on Iris Setosa --------------------*
| The data set contains 50 iris specimens from the species | |
| Iris Setosa with the following four measurements: | |
| SepalLength (sepal length) | |
| SepalWidth (sepal width) | |
| PetalLength (petal length) | |
| PetalWidth (petal width) | |
| Certain values were changed to missing for the analysis. | |
*------------------------------------------------------------*;

data Setosa;
   input SepalLength SepalWidth PetalLength PetalWidth @@;
   label sepallength='Sepal Length in mm.'
           sepalwidth='Sepal Width in mm.'
           petallength='Petal Length in mm.'
           petalwidth='Petal Width in mm.';
datalines;
50 33 14 02 46 34 14 03 46 36 . 02
51 33 17 05 55 35 13 02 48 31 16 02
52 34 14 02 49 36 14 01 44 32 13 02
50 35 16 06 44 30 13 02 47 32 16 02
48 30 14 03 51 38 16 02 48 34 19 02
50 30 16 02 50 32 12 02 43 30 11 .
58 40 12 02 51 38 19 04 49 30 14 02
51 35 14 02 50 34 16 04 46 32 14 02
57 44 15 04 50 36 14 02 54 34 15 04
52 41 15 . 55 42 14 02 49 31 15 02
54 39 17 04 50 34 15 02 44 29 14 02
47 32 13 02 46 31 15 02 51 34 15 02
50 35 13 03 49 31 15 01 54 37 15 02
54 39 13 04 51 35 14 03 48 34 16 02
48 30 14 01 45 23 13 03 57 38 17 03
51 38 15 03 54 34 17 02 51 37 15 04
52 35 15 02 53 37 15 02
;
```

The following statements request a correlation analysis between two sets of variables, the sepal measurements (length and width) and the petal measurements (length and width):

```
ods graphics on;
title 'Fisher (1936) Iris Setosa Data';
proc corr data=Setosa sscp cov plots=matrix;
   var SepalLength SepalWidth;
   with PetalLength PetalWidth;
run;
ods graphics off;
```
The “Simple Statistics” table in Output 2.2.1 displays univariate statistics for variables in the VAR and WITH statements.

**Output 2.2.1 Simple Statistics**

Fisher (1936) Iris Setosa Data

The CORR Procedure

2 With Variables: PetalLength PetalWidth
2 Variables: SepalLength SepalWidth

<table>
<thead>
<tr>
<th>Simple Statistics</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>N</td>
<td>Mean</td>
<td>Std Dev</td>
<td>Sum</td>
</tr>
<tr>
<td>PetalLength</td>
<td>49</td>
<td>14.71429</td>
<td>1.62019</td>
<td>721.0000</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>48</td>
<td>2.52083</td>
<td>1.03121</td>
<td>121.0000</td>
</tr>
<tr>
<td>SepalLength</td>
<td>50</td>
<td>50.06000</td>
<td>3.52490</td>
<td>2503</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>50</td>
<td>34.28000</td>
<td>3.79064</td>
<td>1714</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Statistics</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>PetalLength</td>
<td>11.00000</td>
<td>19.00000</td>
<td>Petal Length in mm.</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>1.00000</td>
<td>6.00000</td>
<td>Petal Width in mm.</td>
</tr>
<tr>
<td>SepalLength</td>
<td>43.00000</td>
<td>58.00000</td>
<td>Sepal Length in mm.</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>23.00000</td>
<td>44.00000</td>
<td>Sepal Width in mm.</td>
</tr>
</tbody>
</table>

When the WITH statement is specified together with the VAR statement, the CORR procedure produces rectangular matrices for statistics such as covariances and correlations. The matrix rows correspond to the WITH variables (PetalLength and PetalWidth), while the matrix columns correspond to the VAR variables (SepalLength and SepalWidth). The CORR procedure uses the WITH variable labels to label the matrix rows.

The SSCP option requests a table of the uncorrected sum-of-squares and crossproducts matrix, and the COV option requests a table of the covariance matrix. The SSCP and COV options also produce a table of the Pearson correlations.

The sum-of-squares and crossproducts statistics for each pair of variables are computed by using observations with nonmissing row and column variable values. The “Sums of Squares and Crossproducts” table in Output 2.2.2 displays the crossproduct, sum of squares for the row variable, and sum of squares for the column variable for each pair of variables.
Output 2.2.2 Sums of Squares and Crossproducts

<table>
<thead>
<tr>
<th></th>
<th>SepalLength</th>
<th>SepalWidth</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PetalLength</td>
<td>36214.00000</td>
<td>24756.00000</td>
<td>10735.00000</td>
<td>10735.00000</td>
<td></td>
</tr>
<tr>
<td>Petal Length in mm.</td>
<td>123793.0000</td>
<td>58164.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PetalWidth</td>
<td>6113.00000</td>
<td>4191.00000</td>
<td>355.00000</td>
<td>355.00000</td>
<td></td>
</tr>
<tr>
<td>Petal Width in mm.</td>
<td>121356.0000</td>
<td>56879.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The variances are computed by using observations with nonmissing row and column variable values. The “Variances and Covariances” table in Output 2.2.3 displays the covariance, variance for the row variable, variance for the column variable, and associated degrees of freedom for each pair of variables.

Output 2.2.3 Variances and Covariances

<table>
<thead>
<tr>
<th></th>
<th>SepalLength</th>
<th>SepalWidth</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PetalLength</td>
<td>1.27083333</td>
<td>1.36309524</td>
<td>2.62500000</td>
<td>2.62500000</td>
<td></td>
</tr>
<tr>
<td>Petal Length in mm.</td>
<td>12.333333</td>
<td>14.605442</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48 48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PetalWidth</td>
<td>0.91134752</td>
<td>1.04831562</td>
<td>1.06338652</td>
<td>1.06338652</td>
<td></td>
</tr>
<tr>
<td>Petal Width in mm.</td>
<td>11.801418</td>
<td>13.627216</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>47 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
When there are missing values in the analysis variables, the “Pearson Correlation Coefficients” table in Output 2.2.4 displays the correlation, the \( p \)-value under the null hypothesis of zero correlation, and the number of observations for each pair of variables. Only the correlation between PedalWidth and SepalLength and the correlation between PedalWidth and SepalWidth are slightly positive.

**Output 2.2.4** Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Sepal Length</th>
<th>Sepal Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>PetalLength</td>
<td>0.22335</td>
<td>0.22014</td>
</tr>
<tr>
<td>Petal Length in mm.</td>
<td>0.1229</td>
<td>0.1285</td>
</tr>
<tr>
<td></td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>0.25726</td>
<td>0.27539</td>
</tr>
<tr>
<td>Petal Width in mm.</td>
<td>0.0775</td>
<td>0.0582</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
</tr>
</tbody>
</table>

When ODS Graphics is enabled, the PLOTS= option displays a scatter matrix plot by default. Output 2.2.5 displays a rectangular scatter plot matrix for the two sets of variables: the VAR variables SepalLength and SepalWidth are listed across the top of the matrix, and the WITH variables PetalLength and PetalWidth are listed down the side of the matrix. As measured in Output 2.2.4, the plot for PedalWidth and SepalLength and the plot for PedalWidth and SepalWidth also show slight positive correlations.
Output 2.2.5 Rectangular Matrix Plot

Note that this graphical display is requested by enabling ODS Graphics and by specifying the PLOTS= option. For more information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS” (SAS/STAT User’s Guide).
Example 2.3: Analysis Using Fisher’s $z$ Transformation

The following statements request Pearson correlation statistics by using Fisher’s $z$ transformation for the data set Fitness:

```latex
proc corr data=Fitness nosimple fisher;
  var weight oxygen runtime;
run;
```

The NOSIMPLE option suppresses the table of univariate descriptive statistics. By default, PROC CORR displays the “Pearson Correlation Coefficients” table in Output 2.3.1.

**Output 2.3.1** Pearson Correlations

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Oxygen</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1.00000</td>
<td>-0.15358</td>
<td>0.20072</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Oxygen</td>
<td>-0.15358</td>
<td>1.00000</td>
<td>-0.86843</td>
</tr>
<tr>
<td></td>
<td>0.4264</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>RunTime</td>
<td>0.20072</td>
<td>-0.86843</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.2965</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>

Using the FISHER option, the CORR procedure displays correlation statistics by using Fisher’s $z$ transformation in Output 2.3.2.
Output 2.3.2 Correlation Statistics Using Fisher’s z Transformation

<table>
<thead>
<tr>
<th>With Variable</th>
<th>Sample Variable</th>
<th>N</th>
<th>Correlation</th>
<th>Fisher’s z</th>
<th>Adjustment</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>Oxygen</td>
<td>29</td>
<td>-0.15358</td>
<td>-0.15480</td>
<td>-0.00274</td>
<td>-0.15090</td>
</tr>
<tr>
<td>Weight</td>
<td>RunTime</td>
<td>29</td>
<td>0.20072</td>
<td>0.20348</td>
<td>0.00358</td>
<td>0.19727</td>
</tr>
<tr>
<td>Oxygen</td>
<td>RunTime</td>
<td>28</td>
<td>-0.86843</td>
<td>-1.32665</td>
<td>-0.01608</td>
<td>-0.86442</td>
</tr>
</tbody>
</table>

The table also displays confidence limits and a p-value for the default null hypothesis $H_0: \rho = \rho_0$. See the section “Fisher’s z Transformation” on page 25 for details on Fisher’s z transformation.

The following statements request one-sided hypothesis tests and confidence limits for the correlations using Fisher’s z transformation:

```
proc corr data=Fitness nosimple nocorr fisher (type=lower);
  var weight oxygen runtime;
run;
```

The NOSIMPLE option suppresses the “Simple Statistics” table, and the NOCORR option suppresses the “Pearson Correlation Coefficients” table.
Output 2.3.3 displays correlation statistics by using Fisher’s $z$ transformation.

**Output 2.3.3** One-Sided Correlation Analysis Using Fisher’s $z$ Transformation

<table>
<thead>
<tr>
<th>With</th>
<th>Sample N</th>
<th>Correlation</th>
<th>Fisher’s $z$</th>
<th>Adjustment</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>Oxygen</td>
<td>29</td>
<td>-0.15358</td>
<td>-0.15480</td>
<td>-0.00274</td>
</tr>
<tr>
<td>Weight</td>
<td>RunTime</td>
<td>29</td>
<td>0.20072</td>
<td>0.20348</td>
<td>0.00358</td>
</tr>
<tr>
<td>Oxygen</td>
<td>RunTime</td>
<td>28</td>
<td>-0.86843</td>
<td>-1.32665</td>
<td>-0.01608</td>
</tr>
</tbody>
</table>

The FISHER(TYPE=LOWER) option requests a lower confidence limit and a $p$-value for the test of the one-sided hypothesis $H_0: \rho \leq 0$ against the alternative hypothesis $H_1: \rho > 0$. Here Fisher’s $z$, the bias adjustment, and the estimate of the correlation are the same as for the two-sided alternative. However, because TYPE=LOWER is specified, only a lower confidence limit is computed for each correlation, and one-sided $p$-values are computed.

**Example 2.4: Applications of Fisher’s $z$ Transformation**

This example illustrates some applications of Fisher’s $z$ transformation. For details, see the section “Fisher’s $z$ Transformation” on page 25.

The following statements simulate independent samples of variables $X$ and $Y$ from a bivariate normal distribution. The first batch of 150 observations is sampled using a known correlation of 0.3, the second batch of 150 observations is sampled using a known correlation of 0.25, and the third batch of 100 observations is sampled using a known correlation of 0.3.

```sas
data Sim (drop=i);
  do i=1 to 400;
    X = rannor(135791);
    Batch = 1 + (i>150) + (i>300);
    if Batch = 1 then Y = 0.3*X + 0.9*rannor(246791);
    if Batch = 2 then Y = 0.25*X + sqrt(.8375)*rannor(246791);
    if Batch = 3 then Y = 0.3*X + 0.9*rannor(246791);
    output;
  end;
runkw;
```
This data set will be used to illustrate the following applications of Fisher’s z transformation:

- testing whether a population correlation is equal to a given value
- testing for equality of two population correlations
- combining correlation estimates from different samples

**Testing Whether a Population Correlation Is Equal to a Given Value \( \rho_0 \)**

You can use the following statements to test the null hypothesis \( H_0: \rho = 0.5 \) against a two-sided alternative \( H_1: \rho \neq 0.5 \). The test is requested with the option FISHER(RHO0=0.5).

```plaintext
title 'Analysis for Batch 1';
proc corr data=Sim (where=(Batch=1)) fisher(rho0=.5);
   var X Y;
run;
```

Output 2.4.1 displays the results based on Fisher’s transformation. The null hypothesis is rejected since the \( p \)-value is less than 0.0001.

**Output 2.4.1** Fisher’s Test for \( H_0 : \rho = \rho_0 \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable</th>
<th>N</th>
<th>Correlation</th>
<th>Fisher's z</th>
<th>Adjustment</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>150</td>
<td>0.22081</td>
<td>0.22451</td>
<td>0.0007410</td>
<td>0.22011</td>
</tr>
</tbody>
</table>

**Testing for Equality of Two Population Correlations**

You can use the following statements to test for equality of two population correlations, \( \rho_1 \) and \( \rho_2 \). Here, the null hypothesis \( H_0: \rho_1 = \rho_2 \) is tested against the alternative \( H_1: \rho_1 \neq \rho_2 \).

```plaintext
ods output FisherPearsonCorr=SimCorr;
title 'Testing Equality of Population Correlations';
proc corr data=Sim (where=(Batch=1 or Batch=2)) fisher;
   var X Y;
   by Batch;
run;
```
The ODS OUTPUT statement saves the “FisherPearsonCorr” table into an output data set in the CORR procedure. The output data set SimCorr contains Fisher’s \( z \) statistics for both batches.

The following statements display (in Figure 2.4.2) the output data set SimCorr:

```sas
proc print data=SimCorr;
run;
```

**Output 2.4.2** Fisher's Correlation Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Batch</th>
<th>Var</th>
<th>Var</th>
<th>NObs</th>
<th>Corr</th>
<th>ZVal</th>
<th>BiasAdj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>X</td>
<td>Y</td>
<td>150</td>
<td>0.22081</td>
<td>0.22451</td>
<td>0.0007410</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>X</td>
<td>Y</td>
<td>150</td>
<td>0.33694</td>
<td>0.35064</td>
<td>0.00113</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>CorrEst</th>
<th>Lcl</th>
<th>Ucl</th>
<th>pValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.22011</td>
<td>0.062034</td>
<td>0.367409</td>
<td>0.0065</td>
</tr>
<tr>
<td>2</td>
<td>0.33594</td>
<td>0.185676</td>
<td>0.470853</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The \( p \)-value for testing \( H_0 \) is derived by treating the difference \( z_1 - z_2 \) as a normal random variable with mean zero and variance \( \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \), where \( z_1 \) and \( z_2 \) are Fisher’s \( z \) transformation of the sample correlations \( r_1 \) and \( r_2 \), respectively, and where \( n_1 \) and \( n_2 \) are the corresponding sample sizes.

The following statements compute the \( p \)-value in Output 2.4.3:

```sas
data SimTest (drop=Batch);
  merge SimCorr (where=(Batch=1) keep=Nobs ZVal Batch rename=(Nobs=n1 ZVal=z1))
    SimCorr (where=(Batch=2) keep=Nobs ZVal Batch rename=(Nobs=n2 ZVal=z2));
  variance = 1/(n1-3) + 1/(n2-3);
  z = (z1 - z2) / sqrt( variance );
  pval = probnorm(z);
  if (pval > 0.5) then pval = 1 - pval;
  pval = 2*pval;
run;

proc print data=SimTest noobs;
run;
```

**Output 2.4.3** Test of Equality of Observed Correlations

<table>
<thead>
<tr>
<th>n1</th>
<th>z1</th>
<th>n2</th>
<th>z2</th>
<th>variance</th>
<th>z</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0.22451</td>
<td>150</td>
<td>0.35064</td>
<td>0.013605</td>
<td>-1.08135</td>
<td>0.27954</td>
</tr>
</tbody>
</table>

In Output 2.4.3, the \( p \)-value of 0.2795 does not provide evidence to reject the null hypothesis that \( \rho_1 = \rho_2 \). The sample sizes \( n_1 = 150 \) and \( n_2 = 150 \) are not large enough to detect the difference \( \rho_1 - \rho_2 = 0.05 \) at a significance level of \( \alpha = 0.05 \).
Combining Correlation Estimates from Different Samples

Assume that sample correlations \( r_1 \) and \( r_2 \) are computed from two independent samples of \( n_1 \) and \( n_2 \) observations, respectively. A combined correlation estimate is given by \( \tilde{r} = \tanh(\tilde{z}) \), where \( \tilde{z} \) is the weighted average of the \( z \) transformations of \( r_1 \) and \( r_2 \):

\[
\tilde{z} = \frac{(n_1 - 3)z_1 + (n_2 - 3)z_2}{n_1 + n_2 - 6}
\]

The following statements compute a combined estimate of \( \rho \) by using Batch 1 and Batch 3:

```r
ods output FisherPearsonCorr=SimCorr2;
proc corr data=Sim (where=(Batch=1 or Batch=3)) fisher;
    var X Y;
    by Batch;
run;

data SimComb (drop=Batch);
    merge SimCorr2 (where=(Batch=1) keep=Nobs ZVal Batch rename=(Nobs=n1 ZVal=z1))
        SimCorr2 (where=(Batch=3) keep=Nobs ZVal Batch rename=(Nobs=n2 ZVal=z2));
    z = ((n1-3)*z1 + (n2-3)*z2) / (n1+n2-6);
    corr = tanh(z);
    var = 1/(n1+n2-6);
    zlcl = z - probit(0.975)\sqrt{var};
    zucl = z + probit(0.975)\sqrt{var};
    lcl= tanh(zlcl);
    ucl= tanh(zucl);
    pval= probnorm( z/sqrt(var));
    if (pval > .5) then pval= 1 - pval;
    pval= 2*pval;
run;

proc print data=SimComb noobs;
    var n1 z1 n2 z2 corr lcl ucl pval;
run;
```

Output 2.4.4 displays the combined estimate of \( \rho \). The table shows that a correlation estimate from the combined samples is \( r = 0.2264 \). The 95% confidence interval is \((0.10453, 0.34156)\), using the variance of the combined estimate. Note that this interval contains the population correlation \( 0.3 \).

**Output 2.4.4 Combined Correlation Estimate**

<table>
<thead>
<tr>
<th>n1</th>
<th>z1</th>
<th>n2</th>
<th>z2</th>
<th>corr</th>
<th>lcl</th>
<th>ucl</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0.22451</td>
<td>100</td>
<td>0.23929</td>
<td>0.22640</td>
<td>0.10453</td>
<td>0.34156</td>
<td>.000319748</td>
</tr>
</tbody>
</table>
Example 2.5: Computing Polyserial Correlations

The following statements create the data set Fitness1. This data set contains an ordinal variable Oxygen that is derived from a continuous measurement of oxygen intake which is not directly observed.

```latex
*----------------- Data on Physical Fitness -----------------*
| These measurements were made on men involved in a physical |
| fitness course at N.C. State University. |
| The variables are Age (years), Weight (kg), |
| Runtime (time to run 1.5 miles in minutes), and |
| Oxygen (an ordinal variable based on oxygen intake, |
| ml per kg body weight per minute) |
| Certain values were changed to missing for the analysis. |
*------------------------------------------------------------*;
```
```
data Fitness1;
    input Age Weight RunTime Oxygen @@;
    datalines;
44  89.47 11.37  8 40  75.07 10.07  9 
44  85.84  8.65 10 42  68.15  8.17  11
38  89.02   .   9 47  77.45 11.63  8 
40  75.98 11.95  9 43  81.19 10.85  9 
44  81.42 13.08  7 38  81.87  8.63 12 
44  73.03 10.13 10 45  87.66 14.03  7 
45  66.45 11.12  8 47  79.15 10.60  9 
54  83.12 10.33 10 49  81.42  8.95  9 
51  69.63 10.95  8 51  77.91 10.00  9 
48  91.63 10.25  9 49  73.37 10.08   .
57  73.37 12.63  7 54  79.38 11.17  9 
52  76.32  9.63  9 50  70.87  8.92 10 
51  67.25 11.08  9 54  91.63 12.88  7 
51  73.71 10.47  9 57  59.08  9.93 10 
49  76.32   .   7 48  61.24 11.50  9 
52  82.78 10.50  9
;```

The following statements compute Pearson correlations and polyserial correlations:

```latex
proc corr data=Fitness1 pearson polyserial;
    with Oxygen;
    var Age Weight RunTime;
run;
```

For the purpose of computing Pearson correlations, the variables in the WITH and VAR statements are treated as continuous variables. For the purpose of computing polyserial correlations, the variables in the WITH statement are treated as ordinal variables by default, and the variables in the VAR statement are treated as continuous variables.
The “Simple Statistics” table in Output 2.5.1 displays univariate descriptive statistics for each analysis variable.

**Output 2.5.1 Simple Statistics**

```
The CORR Procedure

1 With Variables: Oxygen
3 Variables: Age Weight RunTime

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen</td>
<td>29</td>
<td>8.93103</td>
<td>1.16285</td>
<td>9.00000</td>
<td>7.00000</td>
<td>12.00000</td>
</tr>
<tr>
<td>Age</td>
<td>31</td>
<td>47.67742</td>
<td>5.21144</td>
<td>48.00000</td>
<td>38.00000</td>
<td>57.00000</td>
</tr>
<tr>
<td>Weight</td>
<td>31</td>
<td>77.44452</td>
<td>8.32857</td>
<td>77.45000</td>
<td>59.08000</td>
<td>91.63000</td>
</tr>
<tr>
<td>RunTime</td>
<td>29</td>
<td>10.67414</td>
<td>1.39194</td>
<td>10.50000</td>
<td>8.17000</td>
<td>14.03000</td>
</tr>
</tbody>
</table>
```

The “Pearson Correlation Coefficients” table in Output 2.5.2 displays Pearson correlation statistics between Oxygen and the other three variables. The table shows a strong correlation between variables Oxygen and RunTime.

**Output 2.5.2 Pearson Correlation Coefficients**

```
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Weight</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen</td>
<td>-0.25581</td>
<td>-0.22211</td>
<td>-0.85750</td>
</tr>
<tr>
<td></td>
<td>0.1804</td>
<td>0.2469</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>29</td>
<td>28</td>
</tr>
</tbody>
</table>
```
Chapter 2: The CORR Procedure

The “Polyserial Correlations” table in Output 2.5.3 displays polyserial correlation statistics between Oxygen and the three continuous variables. The variable Oxygen is treated as an ordinal variable derived from oxygen intake (the underlying continuous variable), assuming a bivariate normal distribution for oxygen intake and each of the three continuous variables Age, Weight, and RunTime. The CORR procedure provides two tests for a zero polyserial correlation: the Wald test and the likelihood ratio test. The table shows a strong polyserial correlation between RunTime and the underlying continuous variable of Oxygen from both tests.

Output 2.5.3 Polyserial Correlation Coefficients

<table>
<thead>
<tr>
<th>Continuous Variable</th>
<th>Ordinal Variable</th>
<th>N</th>
<th>Correlation</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Oxygen</td>
<td>29</td>
<td>-0.23586</td>
<td>0.18813</td>
<td>1.5717</td>
<td>0.2100</td>
</tr>
<tr>
<td>Weight</td>
<td>Oxygen</td>
<td>29</td>
<td>-0.24514</td>
<td>0.18421</td>
<td>1.7709</td>
<td>0.1833</td>
</tr>
<tr>
<td>RunTime</td>
<td>Oxygen</td>
<td>28</td>
<td>-0.91042</td>
<td>0.04071</td>
<td>500.0345</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Example 2.6: Computing Cronbach’s Coefficient Alpha

The following statements create the data set Fish1 from the Fish data set used in Chapter 93, “The STEPDISC Procedure” (SAS/STAT User’s Guide). The cubic root of the weight (Weight3) is computed as a one-dimensional measure of the size of a fish.

```sas
*------------------- Fish Measurement Data ----------------------*
| The data set contains 35 fish from the species Bream caught in |
| Finland’s lake Laengelmavesi with the following measurements: |
| Weight (in grams) |
| Length3 (length from the nose to the end of its tail, in cm) |
| HtPct (max height, as percentage of Length3) |
| WidthPct (max width, as percentage of Length3) |
*----------------------------------------------------------------*;
data Fish1 (drop=HtPct WidthPct);
title 'Fish Measurement Data';
input Weight Length3 HtPct WidthPct @@;
Weight3= Weight**((1/3));
Height=HtPct*Length3/100;
Width=WidthPct*Length3/100;
datalines;
242.0 30.0 38.4 13.4 290.0 31.2 40.0 13.8
340.0 31.1 39.8 15.1 363.0 33.5 38.0 13.3
```
Example 2.6: Computing Cronbach’s Coefficient Alpha

The following statements request a correlation analysis and compute Cronbach’s coefficient alpha for the variables Weight3, Length3, Height, and Width:

```plaintext
ods graphics on;
   title 'Fish Measurement Data';
   proc corr data=fish1 nomiss alpha plots=matrix;
      var Weight3 Length3 Height Width;
   run;
ods graphics off;
```

The ALPHA option computes Cronbach’s coefficient alpha for the analysis variables.

The “Simple Statistics” table in Output 2.6.1 displays univariate descriptive statistics for each analysis variable.

![Output 2.6.1 Simple Statistics](image)

The table shows the following statistics for each variable:
- **N**: Number of observations
- **Mean**: Average value
- **Std Dev**: Standard deviation
- **Sum**: Total value
- **Minimum**: Smallest value
- **Maximum**: Largest value

For example, the Weight3 variable has an average of 8.44751 with a standard deviation of 0.97574, and ranges from 6.23168 to 10.00000.
The “Pearson Correlation Coefficients” table in Output 2.6.2 displays Pearson correlation statistics for pairs of analysis variables.

**Output 2.6.2  Pearson Correlation Coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Weight3</th>
<th>Length3</th>
<th>Height</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight3</td>
<td>1.00000</td>
<td>0.96523</td>
<td>0.96261</td>
<td>0.92789</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Length3</td>
<td>0.96523</td>
<td>1.00000</td>
<td>0.95492</td>
<td>0.92171</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Height</td>
<td>0.96261</td>
<td>0.95492</td>
<td>1.00000</td>
<td>0.92632</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Width</td>
<td>0.92789</td>
<td>0.92171</td>
<td>0.92632</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Since the data set contains only one species of fish, all the variables are highly correlated. Using the ALPHA option, the CORR procedure computes Cronbach’s coefficient alpha in Output 2.6.3. The Cronbach’s coefficient alpha is a lower bound for the reliability coefficient for the raw variables and the standardized variables. Positive correlation is needed for the alpha coefficient because variables measure a common entity.

**Output 2.6.3  Cronbach's Coefficient Alpha**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.822134</td>
</tr>
<tr>
<td>Standardized</td>
<td>0.985145</td>
</tr>
</tbody>
</table>

Because the variances of some variables vary widely, you should use the standardized score to estimate reliability. The overall standardized Cronbach’s coefficient alpha of 0.985145 provides an acceptable lower bound for the reliability coefficient. This is much greater than the suggested value of 0.70 given by Nunnally and Bernstein (1994).

The standardized alpha coefficient provides information about how each variable reflects the reliability of the scale with standardized variables. If the standardized alpha decreases after removing a variable from the construct, then this variable is strongly correlated with other variables in the scale. On the other hand, if the standardized alpha increases after removing a variable from the construct, then removing this variable from the scale makes the construct more reliable. The “Cronbach Coefficient Alpha with Deleted Variables” table in Output 2.6.4 does not show significant increase or decrease in the standardized alpha coefficients. See the section “Cronbach’s Coefficient Alpha” on page 31 for more information about Cronbach’s alpha.
Output 2.6.4  Cronbach’s Coefficient Alpha with Deleted Variables

<table>
<thead>
<tr>
<th>Deleted Variable</th>
<th>Correlation with Total Alpha</th>
<th>Correlation with Total Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight3</td>
<td>0.975379</td>
<td>0.973464</td>
</tr>
<tr>
<td>Length3</td>
<td>0.967602</td>
<td>0.967177</td>
</tr>
<tr>
<td>Height</td>
<td>0.964715</td>
<td>0.968079</td>
</tr>
<tr>
<td>Width</td>
<td>0.934635</td>
<td>0.937599</td>
</tr>
</tbody>
</table>

Example 2.7: Saving Correlations in an Output Data Set

The following statements compute Pearson correlations:

```
title 'Correlations for a Fitness and Exercise Study';
proc corr data=Fitness nomiss outp=CorrOutp;
  var weight oxygen runtime;
run;
```

The NOMISS option excludes observations with missing values of the VAR statement variables from the analysis—that is, the same set of 28 observations is used to compute the correlation for each pair of variables. The OUTP= option creates an output data set named CorrOutp that contains the Pearson correlation statistics.

“Pearson Correlation Coefficients” table in Output 2.7.1 displays the correlation and the $p$-value under the null hypothesis of zero correlation.

Output 2.7.1  Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Oxygen</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1.00000</td>
<td>-0.18419</td>
<td>0.19505</td>
</tr>
<tr>
<td></td>
<td>0.3481</td>
<td>0.3199</td>
<td></td>
</tr>
<tr>
<td>Oxygen</td>
<td>-0.18419</td>
<td>1.00000</td>
<td>-0.86843</td>
</tr>
<tr>
<td></td>
<td>0.3481</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>RunTime</td>
<td>0.19505</td>
<td>-0.86843</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.3199</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>
The following statements display (in Output 2.7.2) the output data set:

```plaintext
title 'Output Data Set from PROC CORR';
proc print data=CorrOutp noobs;
run;
```

**Output 2.7.2** OUTP= Data Set with Pearson Correlations

<table>
<thead>
<tr>
<th><em>TYPE</em></th>
<th><em>NAME</em></th>
<th>Weight</th>
<th>Oxygen</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td></td>
<td>77.2168</td>
<td>47.1327</td>
<td>10.6954</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>8.4495</td>
<td>5.5535</td>
<td>1.4127</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>28.0000</td>
<td>28.0000</td>
<td>28.0000</td>
</tr>
<tr>
<td>CORR</td>
<td>Weight</td>
<td>1.0000</td>
<td>-0.1842</td>
<td>0.1950</td>
</tr>
<tr>
<td>CORR</td>
<td>Oxygen</td>
<td>-0.1842</td>
<td>1.0000</td>
<td>-0.8684</td>
</tr>
<tr>
<td>CORR</td>
<td>RunTime</td>
<td>0.1950</td>
<td>-0.8684</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

The output data set has the default type CORR and can be used as an input data set for regression or other statistical procedures. For example, the following statements request a regression analysis using CorrOutp, without reading the original data in the REG procedure:

```plaintext
title 'Input Type CORR Data Set from PROC REG';
proc reg data=CorrOutp;
    model runtime= weight oxygen;
run;
```

The following statements generate the same results as the preceding statements:

```plaintext
proc reg data=Fitness;
    model runtime= weight oxygen;
run;
```

---

**Example 2.8: Creating Scatter Plots**

The following statements request a correlation analysis and a scatter plot matrix for the variables in the data set Fish1, which was created in Example 2.6.

```plaintext
ods graphics on;
    title 'Fish Measurement Data';
    proc corr data=fish1 nomiss plots=matrix(histogram);
        var Height Width Length3 Weight3;
    run;
ods graphics off;
```
The “Simple Statistics” table in Output 2.8.1 displays univariate descriptive statistics for analysis variables.

Output 2.8.1 Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>34</td>
<td>15.22</td>
<td>1.98</td>
<td>517.50</td>
<td>11.52</td>
<td>18.96</td>
</tr>
<tr>
<td>Width</td>
<td>34</td>
<td>5.44</td>
<td>0.73</td>
<td>184.89</td>
<td>4.02</td>
<td>6.75</td>
</tr>
<tr>
<td>Length3</td>
<td>34</td>
<td>38.39</td>
<td>4.22</td>
<td>1305</td>
<td>30.00</td>
<td>46.50</td>
</tr>
<tr>
<td>Weight3</td>
<td>34</td>
<td>8.45</td>
<td>0.98</td>
<td>287.22</td>
<td>6.23</td>
<td>10.00</td>
</tr>
</tbody>
</table>

The “Pearson Correlation Coefficients” table in Output 2.8.2 displays Pearson correlation statistics for pairs of analysis variables.

Output 2.8.2 Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Height</th>
<th>Width</th>
<th>Length3</th>
<th>Weight3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>1.00000</td>
<td>0.92632</td>
<td>0.95492</td>
<td>0.96261</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Width</td>
<td>0.92632</td>
<td>1.00000</td>
<td>0.92171</td>
<td>0.92789</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Length3</td>
<td>0.95492</td>
<td>0.92171</td>
<td>1.00000</td>
<td>0.96523</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Weight3</td>
<td>0.96261</td>
<td>0.92789</td>
<td>0.96523</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The variables are highly correlated. For example, the correlation between Height and Width is 0.92632.
The `PLOTS=MATRIX(HISTOGRAM)` option requests a scatter plot matrix for the `VAR` statement variables in Output 2.8.3.

**Output 2.8.3  Scatter Plot Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Height</th>
<th>Width</th>
<th>Length3</th>
<th>Weight3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td><img src="Image" alt="Height" /></td>
<td><img src="Image" alt="Width" /></td>
<td><img src="Image" alt="Length3" /></td>
<td><img src="Image" alt="Weight3" /></td>
</tr>
<tr>
<td>Width</td>
<td><img src="Image" alt="Height" /></td>
<td><img src="Image" alt="Width" /></td>
<td><img src="Image" alt="Length3" /></td>
<td><img src="Image" alt="Weight3" /></td>
</tr>
<tr>
<td>Length3</td>
<td><img src="Image" alt="Height" /></td>
<td><img src="Image" alt="Width" /></td>
<td><img src="Image" alt="Length3" /></td>
<td><img src="Image" alt="Weight3" /></td>
</tr>
<tr>
<td>Weight3</td>
<td><img src="Image" alt="Height" /></td>
<td><img src="Image" alt="Width" /></td>
<td><img src="Image" alt="Length3" /></td>
<td><img src="Image" alt="Weight3" /></td>
</tr>
</tbody>
</table>

Note that this graphical display is requested by enabling ODS Graphics and by specifying the `PLOTS=` option. For more information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS” (SAS/STAT User’s Guide).
To explore the correlation between Height and Width, the following statements display (in Output 2.8.4) a scatter plot with prediction ellipses for the two variables:

```plaintext
ods graphics on;
proc corr data=fish1 nomiss
    plots=scatter(nvar=2 alpha=.20 .30);
    var Height Width Length3 Weight3;
run;
ods graphics off;
```

The PLOTS=SCATTER(NVAR=2) option requests a scatter plot for the first two variables in the VAR list. The ALPHA=.20 .30 suboption requests 80% and 70% prediction ellipses, respectively.

**Output 2.8.4** Scatter Plot with Prediction Ellipses

A prediction ellipse is a region for predicting a new observation from the population, assuming bivariate normality. It also approximates a region that contains a specified percentage of the population. The displayed prediction ellipse is centered at the means \((\bar{x}, \bar{y})\). For further details, see the section “Confidence and Prediction Ellipses” on page 32.
Note that the following statements also display (in Output 2.8.5) a scatter plot for Height and Width:

```
ods graphics on;
proc corr data=fish1
   plots=scatter(alpha=.20 .30);
var Height Width;
run;
ods graphics off;
```

**Output 2.8.5** Scatter Plot with Prediction Ellipses

Output 2.8.5 includes the point (13.9, 5.1), which was excluded from Output 2.8.4 because the observation had a missing value for Weight3. The prediction ellipses in Output 2.8.5 also reflect the inclusion of this observation.

The following statements display (in Output 2.8.6) a scatter plot with confidence ellipses for the mean:

```
ods graphics on;
title 'Fish Measurement Data';
proc corr data=fish1 nomiss
   plots=scatter(ellipse=confidence nvar=2 alpha=.05 .01);
   var Height Width Length3 Weight3;
run;
ods graphics off;
```
The NVAR=2 suboption within the PLOTS= option restricts the number of plots created to the first two variables in the VAR statement, and the ELLIPSE=CONFIDENCE suboption requests confidence ellipses for the mean. The ALPHA=.05 .01 suboption requests 95% and 99% confidence ellipses, respectively.

**Output 2.8.6** Scatter Plot with Confidence Ellipses

The confidence ellipse for the mean is centered at the means \((\bar{x}, \bar{y})\). For further details, see the section “Confidence and Prediction Ellipses” on page 32.

---

**Example 2.9: Computing Partial Correlations**

A partial correlation measures the strength of the linear relationship between two variables, while adjusting for the effect of other variables.

The following statements request a partial correlation analysis of variables Height and Width while adjusting for the variables Length3 and Weight. The latter variables, which are said to be “partialled out” of the analysis, are specified with the PARTIAL statement.
ods graphics on;
title 'Fish Measurement Data';
proc corr data=fish1 plots=scatter(alpha=.20 .30);
  var Height Width;
  partial Length3 Weight3;
run;
ods graphics off;

Output 2.9.1 displays descriptive statistics for all the variables. The partial variance and partial standard deviation for the variables in the VAR statement are also displayed.

**Output 2.9.1 Descriptive Statistics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length3</td>
<td>34</td>
<td>38.385</td>
<td>4.21628</td>
<td>1305</td>
<td>30.000</td>
<td>46.500</td>
</tr>
<tr>
<td>Weight3</td>
<td>34</td>
<td>8.447</td>
<td>0.97574</td>
<td>287.215</td>
<td>6.23168</td>
<td>10.000</td>
</tr>
<tr>
<td>Height</td>
<td>34</td>
<td>15.22</td>
<td>1.98159</td>
<td>517.499</td>
<td>11.520</td>
<td>18.957</td>
</tr>
<tr>
<td>Width</td>
<td>34</td>
<td>5.438</td>
<td>0.72967</td>
<td>184.894</td>
<td>4.020</td>
<td>6.7497</td>
</tr>
</tbody>
</table>

When you specify a PARTIAL statement, observations with missing values are excluded from the analysis. **Output 2.9.2 displays partial correlations for the variables in the VAR statement.**

**Output 2.9.2 Pearson Partial Correlation Coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Partial Variance</th>
<th>Partial Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length3</td>
<td>0.26607</td>
<td>0.51582</td>
</tr>
<tr>
<td>Weight3</td>
<td>0.07315</td>
<td>0.27047</td>
</tr>
</tbody>
</table>

Pearson Partial Correlation Coefficients, N = 34
Prob > |r| under H0: Partial Rho=0

<table>
<thead>
<tr>
<th></th>
<th>Height</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>1.00000</td>
<td>0.25692</td>
</tr>
<tr>
<td></td>
<td>0.25692</td>
<td>0.1558</td>
</tr>
<tr>
<td>Width</td>
<td>0.25692</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.1558</td>
<td></td>
</tr>
</tbody>
</table>
The partial correlation between the variables Height and Width is 0.25692, which is much less than the unpartialed correlation, 0.92632 (in Output 2.9.2). The \( p \)-value for the partial correlation is 0.1558.

The PLOTS=SCATTER option displays (in Output 2.9.3) a scatter plot of the residuals for the variables Height and Width after controlling for the effect of variables Length3 and Weight. The ALPHA=.20 .30 suboption requests 80\% and 70\% prediction ellipses, respectively.

### Output 2.9.3 Partial Residual Scatter Plot

In Output 2.9.3, a standard deviation of Height has roughly the same length on the X axis as a standard deviation of Width on the Y axis. The major axis length is not significantly larger than the minor axis length, indicating a weak partial correlation between Height and Width.
References


Overview: FREQ Procedure

The FREQ procedure produces one-way to $n$-way frequency and contingency (crosstabulation) tables. For two-way tables, PROC FREQ computes tests and measures of association. For $n$-way tables, PROC FREQ provides stratified analysis by computing statistics across, as well as within, strata.

For one-way frequency tables, PROC FREQ computes goodness-of-fit tests for equal proportions or specified null proportions. For one-way tables, PROC FREQ also provides confidence limits and tests for binomial proportions, including tests for noninferiority and equivalence.

For contingency tables, PROC FREQ can compute various statistics to examine the relationships between two classification variables. For some pairs of variables, you might want to examine the existence or strength of any association between the variables. To determine if an association exists, chi-square tests are computed. To estimate the strength of an association, PROC FREQ computes measures of association that tend to be close to zero when there is no association and close to the maximum (or minimum) value when there is perfect association. The statistics for contingency tables include the following:

- chi-square tests and measures
- measures of association
- risks (binomial proportions) and risk differences for $2 \times 2$ tables
- odds ratios and relative risks for $2 \times 2$ tables
- tests for trend
- tests and measures of agreement
- Cochran-Mantel-Haenszel statistics
PROC FREQ computes asymptotic standard errors, confidence intervals, and tests for measures of association and measures of agreement. Exact $p$-values and confidence intervals are available for many test statistics and measures. PROC FREQ also performs analyses that adjust for any stratification variables by computing statistics across, as well as within, strata for $n$-way tables. These statistics include Cochran-Mantel-Haenszel statistics and measures of agreement.

In choosing measures of association to use in analyzing a two-way table, you should consider the study design (which indicates whether the row and column variables are dependent or independent), the measurement scale of the variables (nominal, ordinal, or interval), the type of association that each measure is designed to detect, and any assumptions required for valid interpretation of a measure. You should exercise care in selecting measures that are appropriate for your data.

Similar comments apply to the choice and interpretation of test statistics. For example, the Mantel-Haenszel chi-square statistic requires an ordinal scale for both variables and is designed to detect a linear association. The Pearson chi-square, on the other hand, is appropriate for all variables and can detect any kind of association, but it is less powerful for detecting a linear association because its power is dispersed over a greater number of degrees of freedom (except for $2 \times 2$ tables).

For more information about selecting the appropriate statistical analyses, see Agresti (2007) and Stokes, Davis, and Koch (2012).

Several SAS procedures produce frequency counts; only PROC FREQ computes chi-square tests for one-way to $n$-way tables and measures of association and agreement for contingency tables. Other procedures to consider for counting include the TABULATE and UNIVARIATE procedures. When you want to produce contingency tables and tests of association for sample survey data, use PROC SURVEYFREQ. See Chapter 14, “Introduction to Survey Procedures” (SAS/STAT User’s Guide), for more information. When you want to fit models to categorical data, use a procedure such as CATMOD, GENMOD, GLIMMIX, LOGISTIC, PROBIT, or SURVEYLOGISTIC. See Chapter 8, “Introduction to Categorical Data Analysis Procedures” (SAS/STAT User’s Guide), for more information.

PROC FREQ uses the Output Delivery System (ODS), a SAS subsystem that provides capabilities for displaying and controlling the output from SAS procedures. ODS enables you to convert any of the output from PROC FREQ into a SAS data set. See the section “ODS Table Names” on page 222 for more information.

PROC FREQ uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS” (SAS/STAT User’s Guide). For specific information about the statistical graphics available with the FREQ procedure, see the PLOTS= option in the TABLES statement and the section “ODS Graphics” on page 226.
Getting Started: FREQ Procedure

Frequency Tables and Statistics

The FREQ procedure provides easy access to statistics for testing for association in a crosstabulation table.

In this example, high school students applied for courses in a summer enrichment program; these courses included journalism, art history, statistics, graphic arts, and computer programming. The students accepted were randomly assigned to classes with and without internships in local companies. Table 3.1 contains counts of the students who enrolled in the summer program by gender and whether they were assigned an internship slot.

Table 3.1 Summer Enrichment Data

<table>
<thead>
<tr>
<th>Gender</th>
<th>Internship</th>
<th>Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>boys</td>
<td>yes</td>
<td>35</td>
</tr>
<tr>
<td>boys</td>
<td>no</td>
<td>14</td>
</tr>
<tr>
<td>girls</td>
<td>yes</td>
<td>32</td>
</tr>
<tr>
<td>girls</td>
<td>no</td>
<td>53</td>
</tr>
</tbody>
</table>

The SAS data set SummerSchool is created by inputting the summer enrichment data as cell count data, or providing the frequency count for each combination of variable values. The following DATA step statements create the SAS data set SummerSchool:

```sas
data SummerSchool;
    input Gender $ Internship $ Enrollment $ Count @@;
datalines;
    boys yes yes 35  boys yes no 29
    boys no yes 14  boys no no 27
    girls yes yes 32 girls yes no 10
    girls no yes 53 girls no no 23
;
```

The variable Gender takes the values ‘boys’ or ‘girls,’ the variable Internship takes the values ‘yes’ and ‘no,’ and the variable Enrollment takes the values ‘yes’ and ‘no.’ The variable Count contains the number of students that correspond to each combination of data values. The double at sign (@@) indicates that more than one observation is included on a single data line. In this DATA step, two observations are included on each line.

Researchers are interested in whether there is an association between internship status and summer program enrollment. The Pearson chi-square statistic is an appropriate statistic to assess the association in the corresponding $2 \times 2$ table. The following PROC FREQ statements specify this analysis.

You specify the table for which you want to compute statistics with the TABLES statement. You specify the statistics you want to compute with options after a slash (/) in the TABLES statement.
The ORDER= option controls the order in which variable values are displayed in the rows and columns of the table. By default, the values are arranged according to the alphanumeric order of their unformatted values. If you specify ORDER=DATA, the data are displayed in the same order as they occur in the input data set. Here, because ‘yes’ appears before ‘no’ in the data, ‘yes’ appears first in any table. Other options for controlling order include ORDER=FORMATTED, which orders according to the formatted values, and ORDER=FREQUENCY, which orders by descending frequency count.

In the TABLES statement, Internship*Enrollment specifies a table where the rows are internship status and the columns are program enrollment. The CHISQ option requests chi-square statistics for assessing association between these two variables. Because the input data are in cell count form, the WEIGHT statement is required. The WEIGHT statement names the variable Count, which provides the frequency of each combination of data values.

Figure 3.1 presents the crosstabulation of Internship and Enrollment. In each cell, the values printed under the cell count are the table percentage, row percentage, and column percentage, respectively. For example, in the first cell, 63.21 percent of the students offered courses with internships accepted them and 36.79 percent did not.

**Figure 3.1** Crosstabulation Table

<table>
<thead>
<tr>
<th>Internship</th>
<th>Enrollment</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>67</td>
<td>39</td>
<td>106</td>
<td>30.04</td>
<td>47.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30.04</td>
<td>17.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>63.21</td>
<td>36.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>43.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>67</td>
<td>50</td>
<td>117</td>
<td>30.04</td>
<td>52.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30.04</td>
<td>22.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>57.26</td>
<td>42.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>56.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>89</td>
<td>223</td>
<td>60.09</td>
<td>39.91</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Figure 3.2 displays the statistics produced by the CHISQ option. The Pearson chi-square statistic is labeled ‘Chi-Square’ and has a value of 0.8189 with 1 degree of freedom. The associated $p$-value is 0.3655, which means that there is no significant evidence of an association between internship status and program enrollment. The other chi-square statistics have similar values and are asymptotically equivalent. The other statistics (phi coefficient, contingency coefficient, and Cramér’s V) are measures of association derived from the Pearson chi-square. For Fisher’s exact test, the two-sided $p$-value is 0.4122, which also shows no association between internship status and program enrollment.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>0.8189</td>
<td>0.3655</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>0.8202</td>
<td>0.3651</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>0.5899</td>
<td>0.4425</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>0.8153</td>
<td>0.3666</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.0606</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.0605</td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.0606</td>
<td></td>
</tr>
</tbody>
</table>

Fisher’s Exact Test

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (1,1) Frequency (F)</td>
<td>67</td>
</tr>
<tr>
<td>Left-sided Pr &lt;= F</td>
<td>0.8513</td>
</tr>
<tr>
<td>Right-sided Pr &gt;= F</td>
<td>0.2213</td>
</tr>
<tr>
<td>Table Probability (P)</td>
<td>0.0726</td>
</tr>
<tr>
<td>Two-sided Pr &lt;= P</td>
<td>0.4122</td>
</tr>
</tbody>
</table>

The analysis, so far, has ignored gender. However, it might be of interest to ask whether program enrollment is associated with internship status after adjusting for gender. You can address this question by doing an analysis of a set of tables (in this case, by analyzing the set consisting of one for boys and one for girls). The Cochran-Mantel-Haenszel (CMH) statistic is appropriate for this situation: it addresses whether rows and columns are associated after controlling for the stratification variable. In this case, you would be stratifying by gender.

The PROC FREQ statements for this analysis are very similar to those for the first analysis, except that there is a third variable, Gender, in the TABLES statement. When you cross more than two variables, the two rightmost variables construct the rows and columns of the table, respectively, and the leftmost variables determine the stratification.

The following PROC FREQ statements also request frequency plots for the crosstabulation tables. PROC FREQ produces these plots by using ODS Graphics to create graphs as part of the procedure output. ODS Graphics must be enabled before producing plots. The PLOTS(ONLY)=FREQPLOT option requests frequency plots. The TWOWAY=CLUSTER plot-option specifies a cluster layout for the two-way frequency plots.
ods graphics on;
proc freq data=SummerSchool;
   tables Gender*Internship*Enrollment /
      chisq cmh plots(only)=freqplot(twoway=cluster);
   weight Count;
run;
ods graphics off;

This execution of PROC FREQ first produces two individual crosstabulation tables of Internship by Enrollment: one for boys and one for girls. Frequency plots and chi-square statistics are produced for each individual table. Figure 3.3, Figure 3.4, and Figure 3.5 show the results for boys. Note that the chi-square statistic for boys is significant at the $\alpha = 0.05$ level of significance. Boys offered a course with an internship are more likely to enroll than boys who are not.

Figure 3.4 displays the frequency plot of Internship by Enrollment for boys. By default, frequency plots are displayed as bar charts. You can use PLOTS= options to request dot plots instead of bar charts, to change the orientation of the bars from vertical to horizontal, and to change the scale from frequencies to percents. You can also use PLOTS= options to specify other two-way layouts (stacked, vertical groups, or horizontal groups) and to change the primary grouping from column levels to row levels.

Figure 3.6, Figure 3.7, and Figure 3.8 display the crosstabulation table, frequency plot, and chi-square statistics for girls. You can see that there is no evidence of association between internship offers and program enrollment for girls.

**Figure 3.3** Crosstabulation Table for Boys

<table>
<thead>
<tr>
<th>Internship</th>
<th>Enrollment</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>27</td>
<td>14</td>
<td>41</td>
<td>48.21</td>
<td>53.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25.71</td>
<td>13.33</td>
<td>39.05</td>
<td>25.71</td>
<td>25.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>65.85</td>
<td>34.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>48.21</td>
<td>28.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td>29</td>
<td>35</td>
<td>64</td>
<td>51.79</td>
<td>46.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.62</td>
<td>33.33</td>
<td>60.95</td>
<td>27.62</td>
<td>27.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45.31</td>
<td>54.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>51.79</td>
<td>71.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>56</td>
<td>49</td>
<td>105</td>
<td>53.33</td>
<td>46.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>53.33</td>
<td>46.67</td>
<td>100.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.4 Frequency Plot for Boys

Distribution of Internship by Enrollment Controlling for Gender=boys

Figure 3.5 Chi-Square Statistics for Boys

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>4.2366</td>
<td>0.0396</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>4.2903</td>
<td>0.0383</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>3.4515</td>
<td>0.0632</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>4.1963</td>
<td>0.0405</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.2009</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.1969</td>
<td></td>
</tr>
<tr>
<td>Cramer's V</td>
<td></td>
<td>0.2009</td>
<td></td>
</tr>
</tbody>
</table>

Fisher's Exact Test

<table>
<thead>
<tr>
<th>Cell (1,1) Frequency (F)</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-sided Pr &lt;= F</td>
<td>0.9885</td>
</tr>
<tr>
<td>Right-sided Pr &gt;= F</td>
<td>0.0311</td>
</tr>
<tr>
<td>Table Probability (P)</td>
<td>0.0196</td>
</tr>
<tr>
<td>Two-sided Pr &lt;= P</td>
<td>0.0467</td>
</tr>
</tbody>
</table>
Table 2 of Internship by Enrollment
Controlling for Gender=girls

<table>
<thead>
<tr>
<th>Internship</th>
<th>Enrollment</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>23</td>
<td>19.49</td>
<td>30.26</td>
<td>69.70</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>10</td>
<td>8.47</td>
<td>23.81</td>
<td>30.30</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>33</td>
<td>27.97</td>
<td>27.97</td>
<td>100.00</td>
<td>118</td>
</tr>
</tbody>
</table>

Figure 3.6  Crosstabulation Table for Girls
Chapter 3: The FREQ Procedure

Figure 3.7 Frequency Plot for Girls

Distribution of Internship by Enrollment Controlling for Gender=girls

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>0.5593</td>
<td>0.4546</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>0.5681</td>
<td>0.4510</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>0.2848</td>
<td>0.5936</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>0.5545</td>
<td>0.4565</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.0688</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.0687</td>
<td></td>
</tr>
<tr>
<td>Cramer's V</td>
<td></td>
<td>0.0688</td>
<td></td>
</tr>
</tbody>
</table>

Fisher's Exact Test

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (1,1) Frequency (F)</td>
<td>23</td>
</tr>
<tr>
<td>Left-sided Pr &lt;= F</td>
<td>0.8317</td>
</tr>
<tr>
<td>Right-sided Pr &gt;= F</td>
<td>0.2994</td>
</tr>
<tr>
<td>Table Probability (P)</td>
<td>0.1311</td>
</tr>
<tr>
<td>Two-sided Pr &lt;= P</td>
<td>0.5245</td>
</tr>
</tbody>
</table>
These individual table results demonstrate the occasional problems with combining information into one table and not accounting for information in other variables such as Gender. Figure 3.9 contains the CMH results. There are three summary (CMH) statistics; which one you use depends on whether your rows and/or columns have an order in $r \times c$ tables. However, in the case of $2 \times 2$ tables, ordering does not matter and all three statistics take the same value. The CMH statistic follows the chi-square distribution under the hypothesis of no association, and here, it takes the value 4.0186 with 1 degree of freedom. The associated $p$-value is 0.0450, which indicates a significant association at the $\alpha = 0.05$ level.

Thus, when you adjust for the effect of gender in these data, there is an association between internship and program enrollment. But, if you ignore gender, no association is found. Note that the CMH option also produces other statistics, including estimates and confidence limits for relative risk and odds ratios for $2 \times 2$ tables and the Breslow-Day Test. These results are not displayed here.

**Figure 3.9** Test for the Hypothesis of No Association

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>4.0186</td>
<td>0.0450</td>
</tr>
<tr>
<td>C2</td>
<td>Row Mean Scores Differ</td>
<td>1</td>
<td>4.0186</td>
<td>0.0450</td>
</tr>
<tr>
<td>C3</td>
<td>General Association</td>
<td>1</td>
<td>4.0186</td>
<td>0.0450</td>
</tr>
</tbody>
</table>

**Agreement Study**

Medical researchers are interested in evaluating the efficacy of a new treatment for a skin condition. Dermatologists from participating clinics were trained to conduct the study and to evaluate the condition. After the training, two dermatologists examined patients with the skin condition from a pilot study and rated the same patients. The possible evaluations are terrible, poor, marginal, and clear. Table 3.2 contains the data.

**Table 3.2** Skin Condition Data

<table>
<thead>
<tr>
<th>Dermatologist 1</th>
<th>Dermatologist 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Terrible</td>
</tr>
<tr>
<td>Terrible</td>
<td>10</td>
</tr>
<tr>
<td>Poor</td>
<td>5</td>
</tr>
<tr>
<td>Marginal</td>
<td>2</td>
</tr>
<tr>
<td>Clear</td>
<td>0</td>
</tr>
</tbody>
</table>
The following DATA step statements create the SAS dataset SkinCondition. The dermatologists’ evaluations of the patients are contained in the variables Derm1 and Derm2; the variable Count is the number of patients given a particular pair of ratings.

```sas
data SkinCondition;
  input Derm1 $ Derm2 $ Count;
datalines;
terrible terrible 10
terrible poor 4
terrible marginal 1
terrible clear 0
poor terrible 5
poor poor 10
poor marginal 12
poor clear 2
marginal terrible 2
marginal poor 4
marginal marginal 12
marginal clear 5
clear terrible 0
clear poor 2
clear marginal 6
clear clear 13;
```

The following PROC FREQ statements request an agreement analysis of the skin condition data. In order to evaluate the agreement of the diagnoses (a possible contribution to measurement error in the study), the kappa coefficient is computed.

The TABLES statement requests a crosstabulation of the variables Derm1 and Derm2. The AGREE option in the TABLES statement requests the kappa coefficient, together with its standard error and confidence limits. The KAPPA option in the TEST statement requests a test for the null hypothesis that kappa equals zero, or that the agreement is purely by chance. The NOPRINT option in the TABLES statement suppresses the display of the two-way table. The PLOTS= option requests an agreement plot for the two dermatologists. ODS Graphics must be enabled before producing plots.

```sas
ods graphics on;
proc freq data=SkinCondition order=data;
tables Derm1*Derm2 / agree noprint plots=agreeplot;
test kappa;
weight Count;
run;
ods graphics off;
```

Figure 3.10 and Figure 3.11 show the results. The kappa coefficient has the value 0.3449, which indicates some agreement between the dermatologists, and the hypothesis test confirms that you can reject the null hypothesis of no agreement. This conclusion is further supported by the confidence interval of (0.2030, 0.4868), which suggests that the true kappa is greater than zero. The AGREE option also produces Bowker’s test for symmetry and the weighted kappa coefficient, but that output is not shown here. Figure 3.11 displays the agreement plot for the ratings of the two dermatologists.
Figure 3.10 Agreement Study

The FREQ Procedure

Statistics for Table of Derm1 by Derm2

Simple Kappa Coefficient

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>0.3449</td>
</tr>
<tr>
<td>ASE</td>
<td>0.0724</td>
</tr>
<tr>
<td>95% Lower Conf Limit</td>
<td>0.2030</td>
</tr>
<tr>
<td>95% Upper Conf Limit</td>
<td>0.4868</td>
</tr>
</tbody>
</table>

Test of H0: Kappa = 0

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE under H0</td>
<td>0.0612</td>
</tr>
<tr>
<td>Z</td>
<td>5.6366</td>
</tr>
<tr>
<td>One-sided Pr &gt; Z</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Two-sided Pr &gt;</td>
<td>Z</td>
</tr>
</tbody>
</table>

Figure 3.11 Agreement Plot
Syntax: FREQ Procedure

The following statements are available in the FREQ procedure:

```
PROC FREQ < options > ;
   BY variables ;
   EXACT statistic-options < / computation-options > ;
   OUTPUT < OUT=SAS-data-set > output-options ;
   TABLES requests < / options > ;
   TEST options ;
   WEIGHT variable < / option > ;
```

The PROC FREQ statement is the only required statement for the FREQ procedure. If you specify the following statements, PROC FREQ produces a one-way frequency table for each variable in the most recently created data set.

```
proc freq;
   run;
```

Table 3.3 summarizes the basic functions of the procedure statements. The following sections provide detailed syntax information for the BY, EXACT, OUTPUT, TABLES, TEST, and WEIGHT statements in alphabetical order after the description of the PROC FREQ statement.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BY</td>
<td>Provides separate analyses for each BY group</td>
</tr>
<tr>
<td>EXACT</td>
<td>Requests exact tests</td>
</tr>
<tr>
<td>OUTPUT</td>
<td>Requests an output data set</td>
</tr>
<tr>
<td>TABLES</td>
<td>Specifies tables and requests analyses</td>
</tr>
<tr>
<td>TEST</td>
<td>Requests tests for measures of association and agreement</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>Identifies a weight variable</td>
</tr>
</tbody>
</table>

PROC FREQ Statement

```
PROC FREQ < options > ;
```

The PROC FREQ statement invokes the FREQ procedure. Optionally, it also identifies the input data set. By default, the procedure uses the most recently created SAS data set.

Table 3.4 lists the options available in the PROC FREQ statement. Descriptions of the options follow in alphabetical order.
Table 3.4  PROC FREQ Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPRESS</td>
<td>Begins the next one-way table on the current page</td>
</tr>
<tr>
<td>DATA=</td>
<td>Names the input data set</td>
</tr>
<tr>
<td>FORMCHAR=</td>
<td>Specifies the outline and cell divider characters for crosstabulation tables</td>
</tr>
<tr>
<td>NLEVELS</td>
<td>Displays the number of levels for all TABLES variables</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>Suppresses all displayed output</td>
</tr>
<tr>
<td>ORDER=</td>
<td>Specifies the order for reporting variable values</td>
</tr>
<tr>
<td>PAGE</td>
<td>Displays one table per page</td>
</tr>
</tbody>
</table>

You can specify the following **options** in the PROC FREQ statement.

**COMPRESS**

begins display of the next one-way frequency table on the same page as the preceding one-way table if there is enough space to begin the table. By default, the next one-way table begins on the current page only if the entire table fits on that page. The COMPRESS option is not valid with the PAGE option.

**DATA=SAS-data-set**
	names the **SAS-data-set** to be analyzed by PROC FREQ. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

**FORMCHAR(1,2,7)=‘formchar-string’**

defines the characters to be used for constructing the outlines and dividers for the cells of crosstabulation table displays. The **formchar-string** should be three characters long. The characters are used to draw the vertical separators (1), the horizontal separators (2), and the vertical-horizontal intersections (7). If you do not specify the FORMCHAR= option, PROC FREQ uses FORMCHAR(1,2,7)=’|-+’ by default. Table 3.5 summarizes the formatting characters used by PROC FREQ.

Table 3.5  Formatting Characters Used by PROC FREQ

<table>
<thead>
<tr>
<th>Position</th>
<th>Default</th>
<th>Used to Draw</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Vertical separators</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Horizontal separators</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>Intersections of vertical and horizontal separators</td>
</tr>
</tbody>
</table>

The FORMCHAR= option can specify 20 different SAS formatting characters used to display output; however, PROC FREQ uses only the first, second, and seventh formatting characters. Therefore, the proper specification for PROC FREQ is FORMCHAR(1,2,7)= ‘formchar-string’.

Specifying all blanks for **formchar-string** produces crosstabulation tables with no outlines or dividers— for example, FORMCHAR(1,2,7)=’ ’. You can use any character in **formchar-string**, including hexadecimal characters. If you use hexadecimal characters, you must put an x after the closing quote. For information about which hexadecimal codes to use for which characters, see the documentation for your hardware.

See the CALENDAR, PLOT, and TABULATE procedures in the *Base SAS Procedures Guide* for more information about form characters.
Chapter 3: The FREQ Procedure

NLEVELS displays the “Number of Variable Levels” table, which provides the number of levels for each variable named in the TABLES statements. See the section “Number of Variable Levels Table” on page 214 for details. PROC FREQ determines the variable levels from the formatted variable values, as described in the section “Grouping with Formats” on page 147.

NOPRINT suppresses the display of all output. You can use the NOPRINT option when you only want to create an output data set. See the section “Output Data Sets” on page 211 for information about the output data sets produced by PROC FREQ. Note that the NOPRINT option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, “Using the Output Delivery System” (SAS/STAT User’s Guide).

NOTE: A NOPRINT option is also available in the TABLES statement. It suppresses display of the crosstabulation tables but allows display of the requested statistics.

ORDER=DATA | FORMATTED | FREQ | INTERNAL specifies the order of the variable levels in the frequency and crosstabulation tables, which you request in the TABLES statement.

The ORDER= option can take the following values:

<table>
<thead>
<tr>
<th>Value of ORDER=</th>
<th>Levels Ordered By</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>Order of appearance in the input data set</td>
</tr>
<tr>
<td>FORMATTED</td>
<td>External formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value</td>
</tr>
<tr>
<td>FREQ</td>
<td>Descending frequency count; levels with the most observations come first in the order</td>
</tr>
<tr>
<td>INTERNAL</td>
<td>Unformatted value</td>
</tr>
</tbody>
</table>

By default, ORDER=INTERNAL. The FORMATTED and INTERNAL orders are machine-dependent. The ORDER= option does not apply to missing values, which are always ordered first.

For more information about sort order, see the chapter on the SORT procedure in the Base SAS Procedures Guide and the discussion of BY-group processing in SAS Language Reference: Concepts.

PAGE displays only one table per page. Otherwise, PROC FREQ displays multiple tables per page as space permits. The PAGE option is not valid with the COMPRESS option.

BY Statement

BY variables;

You can specify a BY statement with PROC FREQ to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be
sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the FREQ procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

**EXACT Statement**

```
EXACT statistic-options < / computation-options > ;
```

The EXACT statement requests exact tests and confidence limits for selected statistics. The `statistic-options` identify which statistics to compute, and the `computation-options` specify options for computing exact statistics. See the section “Exact Statistics” on page 206 for details.

**NOTE:** PROC FREQ computes exact tests by using fast and efficient algorithms that are superior to direct enumeration. Exact tests are appropriate when a data set is small, sparse, skewed, or heavily tied. For some large problems, computation of exact tests might require a considerable amount of time and memory. Consider using asymptotic tests for such problems. Alternatively, when asymptotic methods might not be sufficient for such large problems, consider using Monte Carlo estimation of exact $p$-values. You can request Monte Carlo estimation by specifying the MC computation-option in the EXACT statement. See the section “Computational Resources” on page 208 for more information.

**Statistic Options**

The `statistic-options` specify which exact tests and confidence limits to compute. Table 3.6 lists the available `statistic-options` and the exact statistics that are computed. Descriptions of the `statistic-options` follow the table in alphabetical order.

For one-way tables, exact $p$-values are available for binomial proportion tests, the chi-square goodness-of-fit test, and the likelihood ratio chi-square test. Exact (Clopper-Pearson) confidence limits are available for the binomial proportion.

For two-way tables, exact $p$-values are available for the following tests: Pearson chi-square test, likelihood ratio chi-square test, Mantel-Haenszel chi-square test, Fisher’s exact test, Jonckheere-Terpstra test, and Cochran-Armitage test for trend. Exact $p$-values are also available for tests of the following statistics: Pearson correlation coefficient, Spearman correlation coefficient, Kendall’s tau-$b$, Stuart’s tau-$c$, Somers’ $D(C|R)$, Somers’ $D(R|C)$, simple kappa coefficient, and weighted kappa coefficient.

For $2 \times 2$ tables, PROC FREQ provides the exact McNemar’s test, exact confidence limits for the odds ratio, and Barnard’s unconditional exact test for the risk (proportion) difference. PROC FREQ also provides exact unconditional confidence limits for the risk difference and for the relative risk (ratio of proportions). For
stratified $2 \times 2$ tables, PROC FREQ provides Zelen’s exact test for equal odds ratios, exact confidence limits for the common odds ratio, and an exact test for the common odds ratio.

Most of the statistic-option names listed in Table 3.6 are identical to the corresponding option names in the TABLES and OUTPUT statements. You can request exact computations for groups of statistics by using statistic-options that are identical to the TABLES statement options CHISQ, MEASURES, and AGREE. For example, when you specify the CHISQ statistic-option in the EXACT statement, PROC FREQ computes exact $p$-values for the Pearson chi-square, likelihood ratio chi-square, and Mantel-Haenszel chi-square tests for two-way tables. You can request an exact test for an individual statistic by specifying the corresponding statistic-option from the list in Table 3.6.

**Using the EXACT Statement with the TABLES Statement**

You must use a TABLES statement with the EXACT statement. If you use only one TABLES statement, you do not need to specify the same options in both the TABLES and EXACT statements; when you specify a statistic-option in the EXACT statement, PROC FREQ automatically invokes the corresponding TABLES statement option. However, when you use an EXACT statement with multiple TABLES statements, you must specify options in the TABLES statements to request statistics. PROC FREQ then provides exact tests or confidence limits for those statistics that you also specify in the EXACT statement.

<table>
<thead>
<tr>
<th>Statistic Option</th>
<th>Exact Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGREE</td>
<td>McNemar’s test (for $2 \times 2$ tables), simple kappa test, weighted kappa test</td>
</tr>
<tr>
<td>BARNARD</td>
<td>Barnard’s test (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>BINOMIAL</td>
<td>Binomial proportion tests for one-way tables</td>
</tr>
<tr>
<td>CHISQ</td>
<td>Chi-square goodness-of-fit test for one-way tables; Pearson chi-square, likelihood ratio chi-square, and Mantel-Haenszel chi-square tests for two-way tables</td>
</tr>
<tr>
<td>COMOR</td>
<td>Confidence limits for the common odds ratio, common odds ratio test (for $h \times 2 \times 2$ tables)</td>
</tr>
<tr>
<td>EQOR</td>
<td>Zelen’s test for equal odds ratios (for $h \times 2 \times 2$ tables)</td>
</tr>
<tr>
<td>FISHER</td>
<td>Fisher’s exact test</td>
</tr>
<tr>
<td>JT</td>
<td>Jonckheere-Terpstra test</td>
</tr>
<tr>
<td>KAPPA</td>
<td>Test for the simple kappa coefficient</td>
</tr>
<tr>
<td>KENTB</td>
<td>Test for Kendall’s tau-$b$</td>
</tr>
<tr>
<td>LRCHI</td>
<td>Likelihood ratio chi-square test (one-way and two-way tables)</td>
</tr>
<tr>
<td>MCNEM</td>
<td>McNemar’s test (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>MEASURES</td>
<td>Tests for the Pearson correlation and Spearman correlation, confidence limits for the odds ratio (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>MHCHI</td>
<td>Mantel-Haenszel chi-square test</td>
</tr>
<tr>
<td>OR</td>
<td>Confidence limits for the odds ratio (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>PCHI</td>
<td>Pearson chi-square test (one-way and two-way tables)</td>
</tr>
<tr>
<td>PCORR</td>
<td>Test for the Pearson correlation coefficient</td>
</tr>
<tr>
<td>RELRISK</td>
<td>Confidence limits for the relative risk (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>RISKDIFF</td>
<td>Confidence limits for the proportion difference (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>SCORR</td>
<td>Test for the Spearman correlation coefficient</td>
</tr>
<tr>
<td>SMDCR</td>
<td>Test for Somers’ $D(C</td>
</tr>
<tr>
<td>SMDRC</td>
<td>Test for Somers’ $D(R</td>
</tr>
</tbody>
</table>
You can specify the following *statistic-options* in the EXACT statement.

**AGREE**

requests exact tests for the simple and weighted kappa coefficients and McNemar’s exact test. “Tests and Measures of Agreement” on page 192 and “Exact Statistics” on page 206 for details.

The AGREE option in the TABLES statement provides the simple and weighted kappa coefficients (with their standard errors and confidence limits) and the asymptotic McNemar’s test. The AGREE option in the TEST statement provides asymptotic tests for the kappa coefficients.

Kappa coefficients can be computed for square two-way tables, where the number of rows equals the number of columns. For $2 \times 2$ tables, the weighted kappa coefficient equals the simple kappa coefficient, and PROC FREQ displays only the simple kappa coefficient. McNemar’s test is available for $2 \times 2$ tables.

**BARNARD**

requests Barnard’s exact unconditional test for the risk (proportion) difference for $2 \times 2$ tables. See the section “Barnard’s Unconditional Exact Test” on page 184 for details. The RISKDIFF option in the TABLES statement provides risk difference estimates, confidence limits, and asymptotic tests. See the section “Risks and Risk Differences” on page 174 for more information.

**BINOMIAL**

requests exact tests for the binomial proportion (for one-way tables). See the section “Binomial Tests” on page 170 for details. The BINOMIAL option in the TABLES statement provides confidence limits and asymptotic tests for the binomial proportion. See the section “Binomial Proportion” on page 168 for more information.

**CHISQ**

requests exact tests for the Pearson chi-square, likelihood ratio chi-square, and Mantel-Haenszel chi-square for two-way tables. See the section “Chi-Square Tests and Statistics” on page 153 for details. The CHISQ option in the TABLES statement provides asymptotic tests for these statistics.

For one-way tables, the CHISQ option requests the exact chi-square goodness-of-fit test. See the section “Chi-Square Test for One-Way Tables” on page 153 for details.

**COMOR**

requests an exact test and exact confidence limits for the common odds ratio for multiway $2 \times 2$ tables. See the section “Exact Confidence Limits for the Common Odds Ratio” on page 203 for details. The CMH option in the TABLES statement provides Mantel-Haenszel and logit estimates and asymptotic confidence limits for the common odds ratio.

---

**Table 3.6 continued**

<table>
<thead>
<tr>
<th>Statistic Option</th>
<th>Exact Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUTC</td>
<td>Test for Stuart’s tau-c</td>
</tr>
<tr>
<td>TAUC</td>
<td></td>
</tr>
<tr>
<td>TREND</td>
<td>Cochran-Armitage test for trend</td>
</tr>
<tr>
<td>WTKAP</td>
<td>Test for the weighted kappa coefficient</td>
</tr>
<tr>
<td>WTKAPPA</td>
<td></td>
</tr>
</tbody>
</table>
requests Zelen’s exact test for equal odds ratios, which is available for multiway $2 \times 2$ tables. See the section “Zelen’s Exact Test for Equal Odds Ratios” on page 203 for details. The CMH option in the TABLES statement provides the asymptotic Breslow-Day test for homogeneity of odds ratios.

**FISHER**

request Fisher’s exact test. See the sections “Fisher’s Exact Test” on page 157 and “Exact Statistics” on page 206 for details. For $2 \times 2$ tables, the CHISQ option in the TABLES statement provides Fisher’s exact test. For general $R \times C$ tables, Fisher’s exact test is also known as the Freeman-Halton test.

**JT**

requests the exact Jonckheere-Terpstra test. See the sections “Jonckheere-Terpstra Test” on page 190 and “Exact Statistics” on page 206 for details. The JT option in the TABLES statement provides the asymptotic Jonckheere-Terpstra test.

**KAPPA**

requests an exact test for the simple kappa coefficient. See the sections “Simple Kappa Coefficient” on page 193 and “Exact Statistics” on page 206 for details. The AGREE option in the TABLES statement provides the simple kappa estimate, standard error, and confidence limits. The KAPPA option in the TEST statement provides the asymptotic test for the simple kappa coefficient.

Kappa coefficients are defined only for square tables, where the number of rows equals the number of columns. PROC FREQ does not compute kappa coefficients for tables that are not square.

**KENTB**

**TAUB**

requests an exact test for Kendall’s tau-$b$. See the sections “Kendall’s Tau-$b$” on page 160 and “Exact Statistics” on page 206 for details. The MEASURES option in the TABLES statement provides the Kendall’s tau-$b$ estimate and standard error. The KENTB option in the TEST statement provides an asymptotic test for Kendall’s tau-$b$.

**LRCHI**

requests an exact test for the likelihood ratio chi-square for two-way tables. See the sections “Likelihood Ratio Chi-Square Test” on page 156 and “Exact Statistics” on page 206 for details. For one-way tables, the LRCHI option requests an exact likelihood ratio goodness-of-fit test. See the section “Likelihood Ratio Chi-Square Test for One-Way Tables” on page 155 for details.

The CHISQ option in the TABLES statement provides asymptotic likelihood ratio chi-square tests.

**MCNEM**

requests the exact McNemar’s test. See the sections “McNemar’s Test” on page 192 and “Exact Statistics” on page 206 for details. The AGREE option in the TABLES statement provides the asymptotic McNemar’s test.

**MEASURES**

requests exact tests for the Pearson and Spearman correlations. See the sections “Pearson Correlation Coefficient” on page 162, “Spearman Rank Correlation Coefficient” on page 163, and “Exact Statistics” on page 206 for details. The PCORR and SCORR options in the TEST statement provide asymptotic tests for the Pearson and Spearman correlations, respectively.
This option also requests exact confidence limits for the odds ratio for $2 \times 2$ tables. For more information, see the sections “Odds Ratio” on page 186 and “Exact Confidence Limits for the Odds Ratio” on page 187. The MEASURES, OR(CL=), and RELRISK options in the TABLES statement provide asymptotic confidence limits for the odds ratio.

**MHCHI**

requests an exact test for the Mantel-Haenszel chi-square. See the sections “Mantel-Haenszel Chi-Square Test” on page 156 and “Exact Statistics” on page 206 for details. The CHISQ option in the TABLES statement provides the asymptotic Mantel-Haenszel chi-square test.

**OR**

requests exact confidence limits for the odds ratio for $2 \times 2$ tables. For more information, see the sections “Odds Ratio” on page 186 and “Exact Confidence Limits for the Odds Ratio” on page 187. The MEASURES, OR(CL=), and RELRISK options in the TABLES statement provide asymptotic confidence limits for the odds ratio.

You can set the confidence level by specifying the ALPHA= option in the TABLES statement. The default of ALPHA=0.05 produces 95% confidence limits.

**PCHI**

requests an exact test for the Pearson chi-square for two-way tables. See the sections “Pearson Chi-Square Test for Two-Way Tables” on page 154 and “Exact Statistics” on page 206 for details. For one-way tables, the PCHI option requests an exact chi-square goodness-of-fit test. See the section “Chi-Square Test for One-Way Tables” on page 153 for details. The CHISQ option in the TABLES statement provides asymptotic chi-square tests.

**PCORR**

requests an exact test for the Pearson correlation coefficient. See the sections “Pearson Correlation Coefficient” on page 162 and “Exact Statistics” on page 206 for details. The MEASURES option in the TABLES statement provides the Pearson correlation estimate and standard error. The PCORR option in the TEST statement provides an asymptotic test for the Pearson correlation.

**RELRISK < (options) >**

requests exact unconditional confidence limits for the relative risk for $2 \times 2$ tables. PROC FREQ computes the confidence limits by inverting two separate one-sided exact tests (Santner and Snell 1980). By default, this computation uses the unstandardized relative risk as the test statistic. If you specify the RELRISK(METHOD=SCORE) option, the computation uses the score statistic (Chan and Zhang 1999). For more information, see the section “Exact Unconditional Confidence Limits for the Relative Risk” on page 188.

You can set the confidence level by specifying the ALPHA= option in the TABLES statement. The default of ALPHA=0.05 produces 95% confidence limits.

The RELRISK and MEASURES options in the TABLES statement provide asymptotic confidence limits for the relative risk. For more information, see the section “Relative Risks” on page 188.

You can specify the following options:
COLUMN=1 | 2 | BOTH
specifies the table column for which to compute the relative risk. The default is COLUMN=1, which provides exact confidence limits for the column 1 relative risk. If you specify COLUMN=BOTH, PROC FREQ provides exact confidence limits for both column 1 and column 2 relative risks.

METHOD=SCORE
requests exact unconditional confidence limits that are based on the score statistic (Chan and Zhang 1999). For more information, see the section “Exact Unconditional Confidence Limits for the Relative Risk” on page 188. If you do not specify METHOD=SCORE, by default the exact confidence limit computations are based on the unstandardized relative risk.

RISKDIFF < (options) >
requests exact unconditional confidence limits for the risk difference for $2 \times 2$ tables. PROC FREQ computes the confidence limits by inverting two separate one-sided exact tests (Santner and Snell 1980). By default, this computation uses the unstandardized risk difference as the test statistic. If you specify the RISKDIFF(METHOD=SCORE) option, the computation uses the score statistic (Chan and Zhang 1999). See the section “Exact Unconditional Confidence Limits for the Risk Difference” on page 183 for more information.

You can set the confidence level by specifying the ALPHA= option in the TABLES statement. The default of ALPHA=0.05 produces 95% confidence limits.

The RISKDIFF option in the TABLES statement provides asymptotic confidence limits for the risk difference, including Wald, Newcombe, and Miettinen-Nurminen (score) confidence limits. See the section “Risk Difference Confidence Limits” on page 176 for more information.

You can specify the following options:

COLUMN=1 | 2 | BOTH
specifies the table column for which to compute the risk difference. The default is COLUMN=BOTH, which provides exact confidence limits for both column 1 and column 2 risk differences.

METHOD=SCORE
requests exact unconditional confidence limits that are based on the score statistic (Chan and Zhang 1999). See the section “Exact Unconditional Confidence Limits for the Risk Difference” on page 183 for more information. If you do not specify METHOD=SCORE, by default the exact confidence limit computations are based on the unstandardized risk difference.

SCORR
requests an exact test for the Spearman correlation coefficient. See the sections “Spearman Rank Correlation Coefficient” on page 163 and “Exact Statistics” on page 206 for details. The MEASURES option in the TABLES statement provides the Spearman correlation estimate and standard error. The SCORR option in the TEST statement provides an asymptotic test for the Spearman correlation.

SMDCR
requests an exact test for Somers’ $D(C|R)$. See the sections “Somers’ $D$” on page 161 and “Exact Statistics” on page 206 for details. The MEASURES option in the TABLES statement provides Somers’ $D(C|R)$ estimate and the standard error. The SMDCR option in the TEST statement provides an asymptotic test for Somers’ $D(C|R)$.
SMDRC
requests an exact test for Somers’ $D(R|C)$. See the sections “Somers’ $D$” on page 161 and “Exact Statistics” on page 206 for details. The MEASURES option in the TABLES statement provides Somers’ $D(R|C)$ estimate and the standard error. The SMDRC option in the TEST statement requests an asymptotic test for Somers’ $D(C|R)$.

STUTC
requests an exact test for Stuart’s tau-c. See the sections “Stuart’s Tau-c” on page 161 and “Exact Statistics” on page 206 for details. The MEASURES option in the TABLES statement provides Stuart’s tau-c estimate and the standard error. The STUTC option in the TEST statement provides an asymptotic test for Stuart’s tau-c.

TREND
requests the exact Cochran-Armitage test for trend. See the sections “Cochran-Armitage Test for Trend” on page 189 and “Exact Statistics” on page 206 for details. The TREND option in the TABLES statement provides the asymptotic Cochran-Armitage trend test. The trend test is available for tables of dimensions $2 \times C$ or $R \times 2$.

WTKAP
requests an exact test for the weighted kappa coefficient. See the sections “Weighted Kappa Coefficient” on page 194 and “Exact Statistics” on page 206 for details. The AGREE option in the TABLES statement provides the weighted kappa estimate, standard error, and confidence limits. The WTKAP option in the TEST statement provides the asymptotic test for the weighted kappa coefficient.

Kappa coefficients are defined only for square tables, where the number of rows equals the number of columns. PROC FREQ does not compute kappa coefficients for tables that are not square. For $2 \times 2$ tables, the weighted kappa coefficient equals the simple kappa coefficient, and PROC FREQ does not present separate results for the weighted kappa coefficient.

Computation Options
The computation-options specify options for computing exact statistics. You can specify the following computation-options in the EXACT statement after a slash (/).

ALPHA=\alpha
specifies the level of the confidence limits for Monte Carlo $p$-value estimates. The value of $\alpha$ must be between 0 and 1, and the default is 0.01. A confidence level of $\alpha$ produces $100(1 - \alpha)$% confidence limits. The default of ALPHA=.01 produces 99% confidence limits for the Monte Carlo estimates.

The ALPHA= option invokes the MC option.

MAXTIME=value
specifies the maximum clock time (in seconds) that PROC FREQ can use to compute an exact $p$-value. If the procedure does not complete the computation within the specified time, the computation terminates. The MAXTIME= value must be a positive number. This option is available for Monte Carlo estimation of exact $p$-values as well as for direct exact $p$-value computation. For more information, see the section “Computational Resources” on page 208.
MC requests Monte Carlo estimation of exact p-values instead of direct exact p-value computation. Monte Carlo estimation can be useful for large problems that require a considerable amount of time and memory for exact computations but for which asymptotic approximations might not be sufficient. For more information, see the section “Monte Carlo Estimation” on page 209. This option is available for all EXACT statistic-options except the BINOMIAL option and the following options that apply only to $2 \times 2$ or $h \times 2 \times 2$ tables: BARNARD, COMOR, EQOR, MCNEM, OR, RELRISK, and RISKDIFF. PROC FREQ always computes exact tests or confidence limits (not Monte Carlo estimates) for these statistics.

The ALPHA=, N=, and SEED= options invoke the MC option.

MIDP requests exact mid p-values for the exact tests. The exact mid p-value is defined as the exact p-value minus half the exact point probability. For more information, see the section “Definition of p-Values” on page 207.

The MIDP option is available for all EXACT statement statistic-options except the following: BARNARD, EQOR, OR, RELRISK, and RISKDIFF. You cannot specify both the MIDP option and the MC option.

N=n specifies the number of samples for Monte Carlo estimation. The value of n must be a positive integer, and the default is 10,000. Larger values of n produce more precise estimates of exact p-values. Because larger values of n generate more samples, the computation time increases.

The N= option invokes the MC option.

PFORMAT= format-name | EXACT specifies the display format for exact p-values. PROC FREQ applies this format to one- and two-sided exact p-values, exact point probabilities, and exact mid p-values. By default, PROC FREQ displays exact p-values in the PVALUE6.4 format.

You can provide a format-name or you can specify PFORMAT=EXACT to control the format of exact p-values. The value of format-name can be any standard SAS numeric format or a user-defined format. The format length must not exceed 24. For information about formats, see the FORMAT procedure in the Base SAS Procedures Guide and the FORMAT statement and SAS format in SAS Formats and Informats: Reference.

If you specify PFORMAT=EXACT, PROC FREQ uses the 6.4 format to display exact p-values that are greater than or equal to 0.001; the procedure uses the E10.3 format to display values that are between 0.000 and 0.001. This is the format that PROC FREQ uses to display exact p-values in releases before SAS/STAT 12.3. Beginning in SAS/STAT 12.3, by default PROC FREQ uses the PVALUE6.4 format to display exact p-values.

POINT requests exact point probabilities for the exact tests. The exact point probability is the exact probability that the test statistic equals the observed value. For more information, see the section “Definition of p-Values” on page 207.

The POINT option is available for all EXACT statement statistic-options except the following: BARNARD, EQOR, OR, RELRISK, and RISKDIFF. You cannot specify both the POINT option and the MC option.
SEED=number

specifies the initial seed for random number generation for Monte Carlo estimation. The value of the SEED= option must be an integer. If you do not specify the SEED= option or if the SEED= value is negative or zero, PROC FREQ uses the time of day from the computer’s clock to obtain the initial seed. The SEED= option invokes the MC option.

OUTPUT Statement

OUTPUT < OUT=SAS-data-set > output-options ;

The OUTPUT statement creates a SAS data set that contains statistics that are computed by PROC FREQ. Table 3.7 lists the statistics that can be stored in the output data set. You identify which statistics to include by specifying output-options.

You must use a TABLES statement with the OUTPUT statement. The OUTPUT statement stores statistics for only one table request. If you use multiple TABLES statements, the contents of the output data set correspond to the last TABLES statement. If you use multiple table requests in a single TABLES statement, the contents of the output data set correspond to the last table request. Only one OUTPUT statement is allowed in a single invocation of the procedure.

For a one-way or two-way table, the output data set contains one observation that stores the requested statistics for the table. For a multiway table, the output data set contains an observation for each two-way table (stratum) of the multiway crosstabulation. If you request summary statistics for the multiway table, the output data set also contains an observation that stores the across-strata summary statistics. If you use a BY statement, the output data set contains an observation or set of observations for each BY group. For more information about the contents of the output data set, see the section “Contents of the OUTPUT Statement Output Data Set” on page 212.

The output data set that is created by the OUTPUT statement is not the same as the output data set that is created by the OUT= option in the TABLES statement. The OUTPUT statement creates a data set that contains statistics (such as the Pearson chi-square and its p-value), and the OUT= option in the TABLES statement creates a data set that contains frequency table counts and percentages. See the section “Output Data Sets” on page 211 for more information.

As an alternative to the OUTPUT statement, you can use the Output Delivery System (ODS) to store statistics that PROC FREQ computes. ODS can create a SAS data set from any table that PROC FREQ produces. See the section “ODS Table Names” on page 222 for more information.

You can specify the following options in the OUTPUT statement:

OUT=SAS-data-set

specifies the name of the output data set. When you use an OUTPUT statement but do not use the OUT= option, PROC FREQ creates a data set and names it by using the DATA<n> convention.

output-options

specify the statistics to include in the output data set. Table 3.7 lists the output-options that are available in the OUTPUT statement, together with the TABLES statement options that are required to produce the statistics. Descriptions of the output-options follow the table in alphabetical order.
You can specify output-options to request individual statistics, or you can request groups of statistics by using output-options that are identical to the group options in the TABLES statement (for example, the CHISQ, MEASURES, CMH, AGREE, and ALL options).

When you specify an output-option, the output data set includes statistics from the corresponding analysis. In addition to the estimate or test statistic, the output data set includes associated values such as standard errors, confidence limits, p-values, and degrees of freedom. See the section “Contents of the OUTPUT Statement Output Data Set” on page 212 for details.

To store a statistic in the output data set, you must also request computation of that statistic with the appropriate TABLES, EXACT, or TEST statement option. For example, the PCHI output-option includes the Pearson chi-square in the output data set. You must also request computation of the Pearson chi-square by specifying the CHISQ option in the TABLES statement. Or, if you use only one TABLES statement, you can request computation of the Pearson chi-square by specifying the PCHI or CHISQ option in the EXACT statement. Table 3.7 lists the TABLES statement options that are required to produce the OUTPUT data set statistics.

Table 3.7  OUTPUT Statement Output Options

<table>
<thead>
<tr>
<th>Output Option</th>
<th>Output Data Set Statistics</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGRE</td>
<td>McNemar’s test (2 × 2 tables), Bowker’s test, simple and weighted kappas; for multiple strata, overall simple and weighted kappas, tests for equal kappas, and Cochran’s Q (h × 2 × 2 tables)</td>
<td>AGREE</td>
</tr>
<tr>
<td>AJCHI</td>
<td>Continuity-adjusted chi-square (2 × 2 tables)</td>
<td>CHISQ</td>
</tr>
<tr>
<td>ALL</td>
<td>CHISQ, MEASURES, and CMH statistics; N (number of nonmissing observations)</td>
<td>ALL</td>
</tr>
<tr>
<td>BDCHI</td>
<td>Breslow-Day test (h × 2 × 2 tables)</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>BINOMIAL</td>
<td>Binomial statistics (one-way tables)</td>
<td>BINOMIAL</td>
</tr>
<tr>
<td>CHISQ</td>
<td>For one-way tables, goodness-of-fit test; for two-way tables, Pearson, likelihood ratio, continuity-adjusted, and Mantel-Haenszel chi-squares, Fisher’s exact test (2 × 2 tables), phi and contingency coefficients, Cramér’s V</td>
<td>CHISQ</td>
</tr>
<tr>
<td>CMH</td>
<td>Cochran-Mantel-Haenszel (CMH) correlation, row mean scores (ANOVA), and general association statistics; for 2 × 2 tables, logit and Mantel-Haenszel common odds ratios and relative risks, Breslow-Day test</td>
<td>CMH</td>
</tr>
<tr>
<td>CMH1</td>
<td>CMH statistics, except row mean scores (ANOVA) and general association statistics</td>
<td>CMH or CMH1</td>
</tr>
<tr>
<td>CMH2</td>
<td>CMH statistics, except general association statistic</td>
<td>CMH or CMH2</td>
</tr>
<tr>
<td>CMHCOR</td>
<td>CMH correlation statistic</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>CMHGA</td>
<td>CMH general association statistic</td>
<td>CMH</td>
</tr>
<tr>
<td>CMHRMS</td>
<td>CMH row mean scores (ANOVA) statistic</td>
<td>CMH or CMH2</td>
</tr>
<tr>
<td>COCHQ</td>
<td>Cochran’s Q (h × 2 × 2 tables)</td>
<td>AGREE</td>
</tr>
<tr>
<td>CONTGY</td>
<td>Contingency coefficient</td>
<td>CHISQ</td>
</tr>
<tr>
<td>CRAMV</td>
<td>Cramér’s V</td>
<td>CHISQ</td>
</tr>
<tr>
<td>EQKAP</td>
<td>Test for equal simple kappas</td>
<td>AGREE</td>
</tr>
</tbody>
</table>
Table 3.7 continued

<table>
<thead>
<tr>
<th>Output Option</th>
<th>Output Data Set Statistics</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQOR</td>
<td>Zelen’s test for equal odds ratios ( (h \times 2 \times 2 \text{ tables}) )</td>
<td>CMH and EXACT EQOR</td>
</tr>
<tr>
<td>EQWKP</td>
<td>Test for equal weighted kappas</td>
<td>AGREE</td>
</tr>
<tr>
<td>FISHER</td>
<td>Fisher’s exact test</td>
<td>CHISQ or FISHER (^1)</td>
</tr>
<tr>
<td>GAMMA</td>
<td>Gamma</td>
<td>MEASURES</td>
</tr>
<tr>
<td>GS</td>
<td>Gail-Simon test</td>
<td>CMH(GAILSIMON)</td>
</tr>
<tr>
<td>JT</td>
<td>Jonckheere-Terpstra test</td>
<td>JT</td>
</tr>
<tr>
<td>KAPPA</td>
<td>Simple kappa coefficient</td>
<td>AGREE</td>
</tr>
<tr>
<td>KENTB</td>
<td>Kendall’s tau-(b)</td>
<td>MEASURES</td>
</tr>
<tr>
<td>LAMCR</td>
<td>Lambda asymmetric ( (C \mid R) )</td>
<td>MEASURES</td>
</tr>
<tr>
<td>LAMDAS</td>
<td>Lambda symmetric</td>
<td>MEASURES</td>
</tr>
<tr>
<td>LAMRC</td>
<td>Lambda asymmetric ( (R \mid C) )</td>
<td>MEASURES</td>
</tr>
<tr>
<td>LGOR</td>
<td>Logit common odds ratio</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>LGRRRC1</td>
<td>Logit common relative risk, column 1</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>LGRRRC2</td>
<td>Logit common relative risk, column 2</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>LRCHI</td>
<td>Likelihood ratio chi-square</td>
<td>CHISQ</td>
</tr>
<tr>
<td>MCNEM</td>
<td>McNemar’s test ((2 \times 2 \text{ tables}))</td>
<td>AGREE</td>
</tr>
<tr>
<td>MEASURES</td>
<td>Gamma, Kendall’s tau-(b), Stuart’s tau-(c), Somers’ (D(C \mid R)) and (D(R \mid C)), Pearson and Spearman correlations, lambda asymmetric ( (C \mid R) ) and ( (R \mid C) ), lambda symmetric, uncertainty coefficients ( (C \mid R) ) and ( (R \mid C) ), symmetric uncertainty coefficient; odds ratio and relative risks ((2 \times 2 \text{ tables}))</td>
<td>MEASURES</td>
</tr>
<tr>
<td>MHCHI</td>
<td>Mantel-Haenszel chi-square</td>
<td>CHISQ</td>
</tr>
<tr>
<td>MHOR</td>
<td>Mantel-Haenszel common odds ratio</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>MHRRRC1</td>
<td>Mantel-Haenszel common relative risk, column 1</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>MHRRRC2</td>
<td>Mantel-Haenszel common relative risk, column 2</td>
<td>CMH, CMH1, or CMH2</td>
</tr>
<tr>
<td>N</td>
<td>Number of nonmissing observations</td>
<td></td>
</tr>
<tr>
<td>NMISS</td>
<td>Number of missing observations</td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio ((2 \times 2 \text{ tables}))</td>
<td>MEASURES or RELRISK</td>
</tr>
<tr>
<td>PCHI</td>
<td>Chi-square goodness-of-fit test ((\text{one-way tables})), Pearson chi-square ((\text{two-way tables}))</td>
<td>CHISQ</td>
</tr>
<tr>
<td>PCORR</td>
<td>Pearson correlation coefficient</td>
<td>MEASURES</td>
</tr>
<tr>
<td>PHI</td>
<td>Phi coefficient</td>
<td>CHISQ</td>
</tr>
<tr>
<td>PLCORR</td>
<td>Polychoric correlation coefficient</td>
<td>PLCORR</td>
</tr>
<tr>
<td>RDIF1</td>
<td>Column 1 risk difference ((\text{row 1 – row 2}))</td>
<td>RISKDIFF</td>
</tr>
<tr>
<td>RDIF2</td>
<td>Column 2 risk difference ((\text{row 1 – row 2}))</td>
<td>RISKDIFF</td>
</tr>
<tr>
<td>RELRISK</td>
<td>Odds ratio and relative risks ((2 \times 2 \text{ tables}))</td>
<td>MEASURES or RELRISK</td>
</tr>
<tr>
<td>RISKDIFF</td>
<td>Risks and risk differences ((2 \times 2 \text{ tables}))</td>
<td>RELRISK</td>
</tr>
<tr>
<td>RISKDIFF1</td>
<td>Risks and risk difference, column 1</td>
<td>RELRISK</td>
</tr>
<tr>
<td>RISKDIFF2</td>
<td>Risks and risk difference, column 2</td>
<td>RELRISK</td>
</tr>
</tbody>
</table>

\(^1\)CHISQ computes Fisher’s exact test for \(2 \times 2 \text{ tables}\). Use the FISHER option to compute Fisher’s exact test for general \(r \times c\) tables.
Table 3.7  

<table>
<thead>
<tr>
<th>Output Option</th>
<th>Output Data Set Statistics</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRC1</td>
<td>RELRISK1</td>
<td>Relative risk, column 1</td>
</tr>
<tr>
<td>RRC2</td>
<td>RELRISK2</td>
<td>Relative risk, column 2</td>
</tr>
<tr>
<td>RSK1</td>
<td>RISK1</td>
<td>Column 1 overall risk</td>
</tr>
<tr>
<td>RSK11</td>
<td>RISK11</td>
<td>Column 1 risk for row 1</td>
</tr>
<tr>
<td>RSK12</td>
<td>RISK12</td>
<td>Column 2 risk for row 1</td>
</tr>
<tr>
<td>RSK2</td>
<td>RISK2</td>
<td>Column 2 overall risk</td>
</tr>
<tr>
<td>RSK21</td>
<td>RISK21</td>
<td>Column 1 risk for row 2</td>
</tr>
<tr>
<td>RSK22</td>
<td>RISK22</td>
<td>Column 2 risk for row 2</td>
</tr>
<tr>
<td>SCORR</td>
<td>Spearman correlation coefficient</td>
<td>MEASURES</td>
</tr>
<tr>
<td>SMDCR</td>
<td>Somers’ $D(C</td>
<td>R)$</td>
</tr>
<tr>
<td>SMDRC</td>
<td>Somers’ $D(R</td>
<td>C)$</td>
</tr>
<tr>
<td>STUTC</td>
<td>TAUC</td>
<td>Stuart’s tau-c</td>
</tr>
<tr>
<td>TREND</td>
<td>Cochran-Armitage test for trend</td>
<td>TREND</td>
</tr>
<tr>
<td>TSYM</td>
<td>BOWKER</td>
<td>Bowker’s test of symmetry</td>
</tr>
<tr>
<td>U</td>
<td>Symmetric uncertainty coefficient</td>
<td>MEASURES</td>
</tr>
<tr>
<td>UCR</td>
<td>Uncertainty coefficient $(C</td>
<td>R)$</td>
</tr>
<tr>
<td>URC</td>
<td>Uncertainty coefficient $(R</td>
<td>C)$</td>
</tr>
<tr>
<td>WTKAP</td>
<td>WTKAPPA</td>
<td>Weighted kappa coefficient</td>
</tr>
</tbody>
</table>

You can specify the following *output-options* in the OUTPUT statement.

**AGREE**

Includes the following tests and measures of agreement in the output data set: McNemar’s test (for $2 \times 2$ tables), Bowker’s test of symmetry, the simple kappa coefficient, and the weighted kappa coefficient. For multiway tables, the AGREE option also includes the following statistics in the output data set: overall simple and weighted kappa coefficients, tests for equal simple and weighted kappa coefficients, and Cochran’s $Q$ test.

The AGREE option in the TABLES statement requests computation of tests and measures of agreement. See the section “Tests and Measures of Agreement” on page 192 for details about these statistics.

AGREE statistics are computed only for square tables, where the number of rows equals the number of columns. PROC FREQ provides Bowker’s test of symmetry and weighted kappa coefficients only for tables larger than $2 \times 2$. (For $2 \times 2$ tables, Bowker’s test is identical to McNemar’s test, and the weighted kappa coefficient equals the simple kappa coefficient.) Cochran’s $Q$ is available for multiway $2 \times 2$ tables.

**AJCHI**

Includes the continuity-adjusted chi-square in the output data set. The continuity-adjusted chi-square is available for $2 \times 2$ tables and is provided by the CHISQ option in the TABLES statement. See the section “Continuity-Adjusted Chi-Square Test” on page 156 for details.
ALL includes all statistics that are requested by the CHISQ, MEASURES, and CMH output-options in the output data set. ALL also includes the number of nonmissing observations, which you can request individually by specifying the N output-option.

BDCHI includes the Breslow-Day test in the output data set. The Breslow-Day test for homogeneity of odds ratios is computed for multiway $2 \times 2$ tables and is provided by the CMH, CMH1, and CMH2 options in the TABLES statement. See the section “Breslow-Day Test for Homogeneity of the Odds Ratios” on page 202 for details.

BINOMIAL

BIN includes the binomial proportion estimate, confidence limits, and tests in the output data set. The BINOMIAL option in the TABLES statement requests computation of binomial statistics, which are available for one-way tables. See the section “Binomial Proportion” on page 168 for details.

CHISQ includes the following chi-square tests and measures in the output data set for two-way tables: Pearson chi-square, likelihood ratio chi-square, Mantel-Haenszel chi-square, phi coefficient, contingency coefficient, and Cramér’s V. For $2 \times 2$ tables, CHISQ also includes Fisher’s exact test and the continuity-adjusted chi-square in the output data set. See the section “Chi-Square Tests and Statistics” on page 153 for details. For one-way tables, CHISQ includes the chi-square goodness-of-fit test in the output data set. See the section “Chi-Square Test for One-Way Tables” on page 153 for details. The CHISQ option in the TABLES statement requests computation of these statistics.

If you specify the CHISQ(WARN=OUTPUT) option in the TABLES statement, the CHISQ option also includes the variable WARN_PCHI in the output data set. This variable indicates the validity warning for the asymptotic Pearson chi-square test.

CMH includes the following Cochran-Mantel-Haenszel statistics in the output data set: correlation, row mean scores (ANOVA), and general association. For $2 \times 2$ tables, the CMH option also includes the Mantel-Haenszel and logit estimates of the common odds ratio and relative risks. For multiway (stratified) $2 \times 2$ tables, the CMH option includes the Breslow-Day test for homogeneity of odds ratios. The CMH option in the TABLES statement requests computation of these statistics. See the section “Cochran-Mantel-Haenszel Statistics” on page 197 for details.

If you specify the CMH(MF) option in the TABLES statement, the CMH option includes the Mantel-Fleiss analysis in the output data set. The variables MF_CMH and WARN_CMH contain the Mantel-Fleiss criterion and the warning indicator, respectively.

CMH1 includes the CMH statistics in the output data set, with the exception of the row mean scores (ANOVA) statistic and the general association statistic. The CMH1 option in the TABLES statement requests computation of these statistics. See the section “Cochran-Mantel-Haenszel Statistics” on page 197 for details.
CMH2
includes the CMH statistics in the output data set, with the exception of the general association statistic. The CMH2 option in the TABLES statement requests computation of these statistics. See the section “Cochran-Mantel-Haenszel Statistics” on page 197 for details.

CMH2COR
includes the Cochran-Mantel-Haenszel correlation statistic in the output data set. The CMH option in the TABLES statement requests computation of this statistic. See the section “Correlation Statistic” on page 198 for details.

CMH2GA
includes the Cochran-Mantel-Haenszel general association statistic in the output data set. The CMH option in the TABLES statement requests computation of this statistic. See the section “General Association Statistic” on page 199 for details.

CMH2RMS
includes the Cochran-Mantel-Haenszel row mean scores (ANOVA) statistic in the output data set. The CMH option in the TABLES statement requests computation of this statistic. See the section “ANOVA (Row Mean Scores) Statistic” on page 199 for details.

COCHQ
includes Cochran’s Q test in the output data set. The AGREE option in the TABLES statement requests computation of this test, which is available for multiway 2 × 2 tables. See the section “Cochran’s Q Test” on page 196 for details.

CONTGY
includes the contingency coefficient in the output data set. The CHISQ option in the TABLES statement requests computation of the contingency coefficient. See the section “Contingency Coefficient” on page 158 for details.

CRAMV
includes Cramér’s V in the output data set. The CHISQ option in the TABLES statement requests computation of Cramér’s V. See the section “Cramér’s V” on page 158 for details.

EQKAP
includes the test for equal simple kappa coefficients in the output data set. The AGREE option in the TABLES statement requests computation of this test, which is available for multiway, square \((h \times r \times r)\) tables. See the section “Tests for Equal Kappa Coefficients” on page 196 for details.

EQOR
includes Zelen’s exact test for equal odds ratios in the output data set. The EQOR option in the EXACT statement requests computation of this test, which is available for multiway 2 × 2 tables. See the section “Zelen’s Exact Test for Equal Odds Ratios” on page 203 for details.

EQWKP
includes the test for equal weighted kappa coefficients in the output data set. The AGREE option in the TABLES statement requests computation of this test. The test for equal weighted kappas is available for multiway, square \((h \times r \times r)\) tables where \(r > 2\). See the section “Tests for Equal Kappa Coefficients” on page 196 for details.
**FISHER**
includes Fisher’s exact test in the output data set. For $2 \times 2$ tables, the CHISQ option in the TABLES statement provides Fisher’s exact test. For tables larger than $2 \times 2$, the FISHER option in the EXACT statement provides Fisher’s exact test. See the section “Fisher’s Exact Test” on page 157 for details.

**GAMMA**
includes the gamma statistic in the output data set. The MEASURES option in the TABLES statement requests computation of the gamma statistic. See the section “Gamma” on page 160 for details.

**GS GAILSIMON**
includes the Gail-Simon test for qualitative interaction in the output data set. The CMH(GAILSIMON) option in the TABLES statement requests computation of this test. See the section “Gail-Simon Test for Qualitative Interactions” on page 205 for details.

**JT**
includes the Jonckheere-Terpstra test in the output data set. The JT option in the TABLES statement requests the Jonckheere-Terpstra test. See the section “Jonckheere-Terpstra Test” on page 190 for details.

**KAPPA**
includes the simple kappa coefficient in the output data set. The AGREE option in the TABLES statement requests computation of kappa, which is available for square tables (where the number of rows equals the number of columns). For multiway square tables, the KAPPA option also includes the overall kappa coefficient in the output data set. See the sections “Simple Kappa Coefficient” on page 193 and “Overall Kappa Coefficient” on page 196 for details.

**KENTB TAUB**
includes Kendall’s tau-$b$ in the output data set. The MEASURES option in the TABLES statement requests computation of Kendall’s tau-$b$. See the section “Kendall’s Tau-$b$” on page 160 for details.

**LAMCR**
includes the asymmetric lambda $\lambda(C|R)$ in the output data set. The MEASURES option in the TABLES statement requests computation of lambda. See the section “Lambda (Asymmetric)” on page 165 for details.

**LAMDAS**
includes the symmetric lambda in the output data set. The MEASURES option in the TABLES statement requests computation of lambda. See the section “Lambda (Symmetric)” on page 166 for details.

**LAMRC**
includes the asymmetric lambda $\lambda(R|C)$ in the output data set. The MEASURES option in the TABLES statement requests computation of lambda. See the section “Lambda (Asymmetric)” on page 165 for details.
LGOR
includes the logit estimate of the common odds ratio in the output data set. The CMH option in the TABLES statement requests computation of this statistic, which is available for $2 \times 2$ tables. See the section “Adjusted Odds Ratio and Relative Risk Estimates” on page 200 for details.

LGRRC1
includes the logit estimate of the common relative risk (column 1) in the output data set. The CMH option in the TABLES statement requests computation of this statistic, which is available for $2 \times 2$ tables. See the section “Adjusted Odds Ratio and Relative Risk Estimates” on page 200 for details.

LGRRC2
includes the logit estimate of the common relative risk (column 2) in the output data set. The CMH option in the TABLES statement requests computation of this statistic, which is available for $2 \times 2$ tables. See the section “Adjusted Odds Ratio and Relative Risk Estimates” on page 200 for details.

LRCHI
includes the likelihood ratio chi-square in the output data set. The CHISQ option in the TABLES statement requests computation of the likelihood ratio chi-square. See the section “Likelihood Ratio Chi-Square Test” on page 156 for details.

MCNEM
includes McNemar’s test (for $2 \times 2$ tables) in the output data set. The AGREE option in the TABLES statement requests computation of McNemar’s test. See the section “McNemar’s Test” on page 192 for details.

MEASURES
includes the following measures of association in the output data set: gamma, Kendall’s tau-$b$, Stuart’s tau-$c$, Somers’ $D(C \mid R)$, Somers’ $D(R \mid C)$, Pearson and Spearman correlation coefficients, lambda (symmetric and asymmetric), and uncertainty coefficients (symmetric and asymmetric). For $2 \times 2$ tables, the MEASURES option also includes the odds ratio, column 1 relative risk, and column 2 relative risk. The MEASURES option in the TABLES statement requests computation of these statistics. See the section “Measures of Association” on page 158 for details.

MHCHI
includes the Mantel-Haenszel chi-square in the output data set. The CHISQ option in the TABLES statement requests computation of the Mantel-Haenszel chi-square. See the section “Mantel-Haenszel Chi-Square Test” on page 156 for details.

MHOR
COMOR
includes the Mantel-Haenszel estimate of the common odds ratio in the output data set. The CMH option in the TABLES statement requests computation of this statistic, which is available for $2 \times 2$ tables. See the section “Adjusted Odds Ratio and Relative Risk Estimates” on page 200 for details.

MHRRC1
includes the Mantel-Haenszel estimate of the common relative risk (column 1) in the output data set. The CMH option in the TABLES statement requests computation of this statistic, which is available for $2 \times 2$ tables. See the section “Adjusted Odds Ratio and Relative Risk Estimates” on page 200 for details.
includes the Mantel-Haenszel estimate of the common relative risk (column 2) in the output data set. The CMH option in the TABLES statement requests computation of this statistic, which is available for $2 \times 2$ tables. See the section “Adjusted Odds Ratio and Relative Risk Estimates” on page 200 for details.

includes the number of nonmissing observations in the output data set.

includes the number of missing observations in the output data set. See the section “Missing Values” on page 148 for details about how PROC FREQ handles missing values.

includes the odds ratio (for $2 \times 2$ tables) in the output data set. The MEASURES, OR(CL=), and RELRISK options in the TABLES statement request this statistic. For more information, see the section “Odds Ratio” on page 186.

includes the Pearson chi-square in the output data set for two-way tables. See the section “Pearson Chi-Square Test for Two-Way Tables” on page 154 for details. For one-way tables, the PCHI option includes the chi-square goodness-of-fit test in the output data set. See the section “Chi-Square Test for One-Way Tables” on page 153 for details. The CHISQ option in the TABLES statement requests computation of these statistics.

If you specify the CHISQ(WARN=OUTPUT) option in the TABLES statement, the PCHI option also includes the variable WARN_PCHI in the output data set. This variable indicates the validity warning for the asymptotic Pearson chi-square test.

includes the Pearson correlation coefficient in the output data set. The MEASURES option in the TABLES statement requests computation of the Pearson correlation. See the section “Pearson Correlation Coefficient” on page 162 for details.

includes the phi coefficient in the output data set. The CHISQ option in the TABLES statement requests computation of the phi coefficient. See the section “Phi Coefficient” on page 158 for details.

includes the polychoric correlation coefficient in the output data set. For $2 \times 2$ tables, this statistic is known as the tetrachoric correlation coefficient. The PLCORR option in the TABLES statement requests computation of the polychoric correlation. See the section “Polychoric Correlation” on page 164 for details.

includes the column 1 risk difference (row 1 – row 2) in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for $2 \times 2$ tables. See the section “Risks and Risk Differences” on page 174 for details.
RDIF2 includes the column 2 risk difference (row 1 – row 2) in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for 2 × 2 tables. See the section “Risks and Risk Differences” on page 174 for details.

RELRISK includes the column 1 and column 2 relative risks (for 2 × 2 tables) in the output data set. The MEASURES and RELRISK options in the TABLES statement request these statistics. For more information, see the section “Relative Risks” on page 188.

RISKDIFF includes risks (binomial proportions) and risk differences for 2 × 2 tables in the output data set. These statistics include the row 1 risk, row 2 risk, total (overall) risk, and risk difference (row 1 – row 2) for column 1 and column 2. The RISKDIFF option in the TABLES statement requests computation of these statistics. See the section “Risks and Risk Differences” on page 174 for details.

RISKDIFF1 includes column 1 risks (binomial proportions) and risk differences for 2 × 2 tables in the output data set. These statistics include the row 1 risk, row 2 risk, total (overall) risk, and risk difference (row 1 – row 2). The RISKDIFF option in the TABLES statement requests computation of these statistics. See the section “Risks and Risk Differences” on page 174 for details.

RISKDIFF2 includes column 2 risks (binomial proportions) and risk differences for 2 × 2 tables in the output data set. These statistics include the row 1 risk, row 2 risk, total (overall) risk, and risk difference (row 1 – row 2). The RISKDIFF option in the TABLES statement requests computation of these statistics. See the section “Risks and Risk Differences” on page 174 for details.

RRC1

RELRISK1 includes the column 1 relative risk in the output data set. The MEASURES and RELRISK options in the TABLES statement request relative risks, which are available for 2 × 2 tables. See the section “Odds Ratio and Relative Risks for 2 × 2 Tables” on page 186 for details.

RRC2

RELRISK2 includes the column 2 relative risk in the output data set. The MEASURES and RELRISK options in the TABLES statement request relative risks, which are available for 2 × 2 tables. See the section “Odds Ratio and Relative Risks for 2 × 2 Tables” on page 186 for details.

RSK1

RISK1 includes the overall column 1 risk in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for 2 × 2 tables. See the section “Risks and Risk Differences” on page 174 for details.

RSK11
RISK11
includes the column 1 risk for row 1 in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for $2 \times 2$ tables. See the section “Risks and Risk Differences” on page 174 for details.

RSK12
includes the column 2 risk for row 1 in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for $2 \times 2$ tables. See the section “Risks and Risk Differences” on page 174 for details.

RISK2
includes the overall column 2 risk in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences. See the section “Risks and Risk Differences” on page 174 for details.

RSK21
includes the column 1 risk for row 2 in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for $2 \times 2$ tables. See the section “Risks and Risk Differences” on page 174 for details.

RSK22
includes the column 2 risk for row 2 in the output data set. The RISKDIFF option in the TABLES statement requests computation of risks and risk differences, which are available for $2 \times 2$ tables. See the section “Risks and Risk Differences” on page 174 for details.

SCORR
includes the Spearman correlation coefficient in the output data set. The MEASURES option in the TABLES statement requests computation of the Spearman correlation. See the section “Spearman Rank Correlation Coefficient” on page 163 for details.

SMDCR
includes Somers’ $D(C|R)$ in the output data set. The MEASURES option in the TABLES statement requests computation of Somers’ $D$. See the section “Somers’ $D$” on page 161 for details.

SMDRC
includes Somers’ $D(R|C)$ in the output data set. The MEASURES option in the TABLES statement requests computation of Somers’ $D$. See the section “Somers’ $D$” on page 161 for details.

STUTC
includes Stuart’s tau-c in the output data set. The MEASURES option in the TABLES statement requests computation of tau-c. See the section “Stuart’s Tau-c” on page 161 for details.
TREND
includes the Cochran-Armitage test for trend in the output data set. The TREND option in the TABLES
statement requests computation of the trend test. This test is available for tables of dimension $2 \times C$ or
$R \times 2$. See the section “Cochran-Armitage Test for Trend” on page 189 for details.

TSYMM
BOWKER
includes Bowker’s test of symmetry in the output data set. The AGREE option in the TABLES
statement requests computation of Bowker’s test. See the section “Bowker’s Test of Symmetry” on
page 193 for details.

U
includes the uncertainty coefficient (symmetric) in the output data set. The MEASURES option in the
TABLES statement requests computation of the uncertainty coefficient. See the section “Uncertainty
Coefficient (Symmetric)” on page 167 for details.

UCR
includes the asymmetric uncertainty coefficient $U(C|R)$ in the output data set. The MEASURES
option in the TABLES statement requests computation of the uncertainty coefficient. See the section
“Uncertainty Coefficients (Asymmetric)” on page 167 for details.

URC
includes the asymmetric uncertainty coefficient $U(R|C)$ in the output data set. The MEASURES
option in the TABLES statement requests computation of the uncertainty coefficient. See the section
“Uncertainty Coefficients (Asymmetric)” on page 167 for details.

WTKAP
WTKAPPA
includes the weighted kappa coefficient in the output data set. The AGREE option in the TABLES
statement requests computation of weighted kappa, which is available for square tables larger than
$2 \times 2$. For multiway tables, the WTKAP option also includes the overall weighted kappa coefficient in
the output data set. See the sections “Weighted Kappa Coefficient” on page 194 and “Overall Kappa
Coefficient” on page 196 for details.

### TABLES Statement

```
TABLES requests < / options > ;
```

The TABLES statement requests one-way to $n$-way frequency and crosstabulation tables and statistics for
those tables.

If you omit the TABLES statement, PROC FREQ generates one-way frequency tables for all data set variables
that are not listed in the other statements.

The following argument is required in the TABLES statement.

requests
specify the frequency and crosstabulation tables to produce. A request is composed of one variable
name or several variable names separated by asterisks. To request a one-way frequency table, use a
single variable. To request a two-way crosstabulation table, use an asterisk between two variables. To request a multiway table (an \(n\)-way table, where \(n>2\)), separate the desired variables with asterisks. The unique values of these variables form the rows, columns, and strata of the table. You can include up to 50 variables in a single multiway table request.

For two-way to multiway tables, the values of the last variable form the crosstabulation table columns, while the values of the next-to-last variable form the rows. Each level (or combination of levels) of the other variables forms one stratum. PROC FREQ produces a separate crosstabulation table for each stratum. For example, a specification of \(A^*B^*C^*D\) in a TABLES statement produces \(k\) tables, where \(k\) is the number of different combinations of values for \(A\) and \(B\). Each table lists the values for \(C\) down the side and the values for \(D\) across the top.

You can use multiple TABLES statements in the PROC FREQ step. PROC FREQ builds all the table requests in one pass of the data, so that there is essentially no loss of efficiency. You can also specify any number of table requests in a single TABLES statement. To specify multiple table requests quickly, use a grouping syntax by placing parentheses around several variables and joining other variables or variable combinations. For example, the statements shown in Table 3.8 illustrate grouping syntax.

<table>
<thead>
<tr>
<th>TABLES Request</th>
<th>Equivalent to</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A^*B)</td>
<td>(A^*B)</td>
</tr>
<tr>
<td>((A B)^*C)</td>
<td>(A^*C)</td>
</tr>
<tr>
<td>((A B)^*C)</td>
<td>(A^*C)</td>
</tr>
<tr>
<td>((A B)^*C)</td>
<td>(A^*C)</td>
</tr>
<tr>
<td>((A B)^*C)</td>
<td>(A^*C)</td>
</tr>
<tr>
<td>((A B)^*C)</td>
<td>(A^*C)</td>
</tr>
</tbody>
</table>

The TABLES statement variables are one or more variables from the DATA= input data set. These variables can be either character or numeric, but the procedure treats them as categorical variables. PROC FREQ uses the formatted values of the TABLES variable to determine the categorical variable levels. So if you assign a format to a variable with a FORMAT statement, PROC FREQ formats the values before dividing observations into the levels of a frequency or crosstabulation table. See the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

If you use PROC FORMAT to create a user-written format that combines missing and nonmissing values into one category, PROC FREQ treats the entire category of formatted values as missing. See the discussion in the section “Grouping with Formats” on page 147 for more information.

By default, the frequency or crosstabulation table lists the values of both character and numeric variables in ascending order based on internal (unformatted) variable values. You can change the order of the values in the table by specifying the ORDER= option in the PROC FREQ statement. To list the values in ascending order by formatted value, use ORDER=FORMATTED.

**Without Options**

If you request a one-way frequency table for a variable without specifying options, PROC FREQ produces frequencies, cumulative frequencies, percentages of the total frequency, and cumulative percentages for each value of the variable. If you request a two-way or an \(n\)-way crosstabulation table without specifying
options, PROC FREQ produces crosstabulation tables that include cell frequencies, cell percentages of the total frequency, cell percentages of row frequencies, and cell percentages of column frequencies. The procedure excludes observations with missing values from the table but displays the total frequency of missing observations below each table.

**Options**

Table 3.9 lists the options available in the TABLES statement. Descriptions of the options follow in alphabetical order.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control Statistical Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>AGREE</td>
<td>Requests tests and measures of classification agreement</td>
</tr>
<tr>
<td>ALL</td>
<td>Requests tests and measures of association produced by the CHISQ, MEASURES, and CMH options</td>
</tr>
<tr>
<td>ALPHA=</td>
<td>Sets confidence level for confidence limits</td>
</tr>
<tr>
<td>BINOMIAL</td>
<td>BIN</td>
</tr>
<tr>
<td>CHISQ</td>
<td>Requests chi-square tests and measures based on chi-square</td>
</tr>
<tr>
<td>CL</td>
<td>Requests confidence limits for MEASURES statistics</td>
</tr>
<tr>
<td>CMH</td>
<td>Requests all Cochran-Mantel-Haenszel statistics</td>
</tr>
<tr>
<td>CMH1</td>
<td>Requests CMH correlation statistic, adjusted odds ratios, and adjusted relative risks</td>
</tr>
<tr>
<td>CMH2</td>
<td>Requests CMH correlation and row mean scores (ANOVA) statistics, adjusted odds ratios, and adjusted relative risks</td>
</tr>
<tr>
<td>FISHER</td>
<td>Requests Fisher’s exact test for tables larger than $2 \times 2$</td>
</tr>
<tr>
<td>GAILSIMON</td>
<td>Requests Gail-Simon test for qualitative interactions</td>
</tr>
<tr>
<td>JT</td>
<td>Requests Jonckheere-Terpstra test</td>
</tr>
<tr>
<td>MEASURES</td>
<td>Requests measures of association</td>
</tr>
<tr>
<td>MISSING</td>
<td>Treats missing values as nonmissing</td>
</tr>
<tr>
<td>PLCORR</td>
<td>Requests polychoric correlation</td>
</tr>
<tr>
<td>RELRISK</td>
<td>Requests odds ratio and relative risks for $2 \times 2$ tables</td>
</tr>
<tr>
<td>RISKDIFF</td>
<td>Requests risks and risk differences for $2 \times 2$ tables</td>
</tr>
<tr>
<td>SCORES=</td>
<td>Specifies type of row and column scores</td>
</tr>
<tr>
<td>TREND</td>
<td>Requests Cochran-Armitage test for trend</td>
</tr>
<tr>
<td><strong>Control Additional Table Information</strong></td>
<td></td>
</tr>
<tr>
<td>CELLCI2</td>
<td>Displays cell contributions to the Pearson chi-square statistic</td>
</tr>
<tr>
<td>CUMCOL</td>
<td>Displays cumulative column percentages</td>
</tr>
<tr>
<td>DEVIATION</td>
<td>Displays deviations of cell frequencies from expected values</td>
</tr>
<tr>
<td>EXPECTED</td>
<td>Displays expected cell frequencies</td>
</tr>
<tr>
<td>MISSPRINT</td>
<td>Displays missing value frequencies</td>
</tr>
<tr>
<td>PEARSONRES</td>
<td>Displays Pearson residuals in the CROSSLIST table</td>
</tr>
<tr>
<td>PRINTKWTS</td>
<td>Displays kappa coefficient weights</td>
</tr>
<tr>
<td>SCOROUT</td>
<td>Displays row and column scores</td>
</tr>
<tr>
<td>SPARSE</td>
<td>Includes all possible combinations of variable levels in the LIST table and OUT= data set</td>
</tr>
<tr>
<td>STDRES</td>
<td>Displays standardized residuals in the CROSSLIST table</td>
</tr>
</tbody>
</table>
Table 3.9  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTPCT</td>
<td>Displays percentages of total frequency for ( n )-way tables (( n \geq 2 ))</td>
</tr>
</tbody>
</table>

Control Displayed Output

- CONTENTS= Specifies contents label for crosstabulation tables
- CROSSLIST Displays crosstabulation tables in ODS column format
- FORMAT= Formats frequencies in crosstabulation tables
- LIST Displays two-way to \( n \)-way tables in list format
- MAXLEVELS= Specifies maximum number of levels to display in one-way tables
- NOCOL Suppresses display of column percentages
- NOCUM Suppresses display of cumulative frequencies and percentages
- NOFREQ Suppresses display of frequencies
- NOPERCENT Suppresses display of percentages
- NOPRINT Suppresses display of crosstabulation tables but displays statistics
- NOROW Suppresses display of row percentages
- NOSPARSE Suppresses zero frequency levels in the CROSSLIST table, LIST table, and OUT= data set
- NOWARN Suppresses log warning message for the chi-square test

Produce Statistical Graphics

- PLOTS= Requests plots from ODS Graphics

Create an Output Data Set

- OUT= Names an output data set to contain frequency counts
- OUTCUM Includes cumulative frequencies and percentages in the output data set for one-way tables
- OUTEXPECT Includes expected frequencies in the output data set
- OUTFPCT Includes row, column, and two-way table percentages in the output data set

You can specify the following options in a TABLES statement.

**AGREE < (options)>**

requests tests and measures of classification agreement for square tables. This option provides the simple and weighted kappa coefficients along with their standard errors and confidence limits. For multiway tables, the AGREE option also produces the overall simple and weighted kappa coefficients (along with their standard errors and confidence limits) and tests for equal kappas among strata. For \( 2 \times 2 \) tables, this option provides McNemar’s test; for square tables that have more than two response categories (levels), this option provides Bowker’s test of symmetry. For multiway tables that have two response categories, the AGREE option also produces Cochran’s \( Q \) test. For more information, see the section “Tests and Measures of Agreement” on page 192.

Measures of agreement can be computed only for square tables, where the number of rows equals the number of columns. If your table is not square because of observations that have zero weights, you can specify the ZEROS option in the WEIGHT statement to include these observations in the analysis. For more information, see the section “Tables with Zero Rows and Columns” on page 196.
You can set the level for agreement confidence limits by specifying the ALPHA= option in the TABLES statement. The default of ALPHA=0.05 produces 95% confidence limits.

You can specify the TEST statement to request asymptotic tests for the simple and weighted kappa coefficients. You can specify the EXACT statement to request McNemar’s exact test (for $2 \times 2$ tables) and exact tests for the simple and weighted kappa coefficients. For more information, see the section “Exact Statistics” on page 206.

The weighted kappa coefficient is computed by using agreement weights that reflect the relative agreement between pairs of variable levels. To specify the type of agreement weights and to display the agreement weights, you can specify the following options:

**PRINTKWTS**

displays the agreement weights that PROC FREQ uses to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. By default, PROC FREQ uses the Cicchetti-Allison form of agreement weights. If you specify the WT=FC option, the procedure uses the Fleiss-Cohen form of agreement weights. For more information, see the section “Weighted Kappa Coefficient” on page 194.

**WT=FC**

requests Fleiss-Cohen agreement weights in the weighted kappa computation. By default, PROC FREQ uses Cicchetti-Allison agreement weights to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. For more information, see the section “Weighted Kappa Coefficient” on page 194.

**ALL**

requests all tests and measures that are produced by the CHISQ, MEASURES, and CMH options. You can control the number of CMH statistics to compute by specifying the CMH1 or CMH2 option.

**ALPHA=**

specifies the level of confidence limits. The value of $\alpha$ must be between 0 and 1, and the default is 0.05. A confidence level of $\alpha$ produces $100(1 - \alpha)$% confidence limits. The default of ALPHA=0.05 produces 95% confidence limits.

ALPHA= applies to confidence limits that are requested by TABLES statement options. There is a separate ALPHA= option in the EXACT statement that sets the level of confidence limits for Monte Carlo estimates of exact $p$-values, which you request by specifying the MC option in the EXACT statement.

**BINOMIAL < (binomial-options) >**

**BIN < (binomial-options) >**

requests the binomial proportion for one-way tables. When you specify the BINOMIAL option, by default PROC FREQ also provides the asymptotic standard error, asymptotic Wald and exact (Clopper-Pearson) confidence limits, and the asymptotic equality test for the binomial proportion.

You can specify binomial-options in parentheses after the BINOMIAL option. The LEVEL= binomial-option identifies the variable level for which to compute the proportion. If you do not specify LEVEL=, PROC FREQ computes the proportion for the first level that appears in the output. The P= binomial-option specifies the null proportion for the binomial tests. If you do not specify P=, PROC FREQ uses P=0.5 by default.
You can also specify `binomial-options` to request additional tests and confidence limits for the binomial proportion. The `EQUIV`, `NONINF`, and `SUP` `binomial-options` request tests of equivalence, noninferiority, and superiority, respectively. The `CL=` `binomial-option` requests confidence limits for the binomial proportion. Table 3.10 summarizes the `binomial-options`.

Available confidence limits for the binomial proportion include Agresti-Coull, exact (Clopper-Pearson), Jeffreys, Wald, and Wilson (score) confidence limits. You can specify more than one type of binomial confidence limits in the same analysis. If you do not specify any confidence limit requests, PROC FREQ computes Wald asymptotic confidence limits and exact (Clopper-Pearson) confidence limits by default. The `ALPHA=` option determines the confidence level, and the default of `ALPHA=0.05` produces 95% confidence limits for the binomial proportion.

As part of the noninferiority, superiority, and equivalence analyses, PROC FREQ provides test-based confidence limits that have a confidence coefficient of $100(1 - 2\alpha)$% (Schuirmann 1999). The `ALPHA=` option determines the confidence level, and the default of `ALPHA=0.05` produces 90% confidence limits. See the sections “Noninferiority Test” on page 171 and “Equivalence Test” on page 173 for details.

To request exact tests for the binomial proportion, specify the `BINOMIAL` option in the `EXACT` statement. PROC FREQ then computes exact $p$-values for all binomial tests that you request with `binomial-options`, which can include tests of noninferiority, superiority, and equivalence, in addition to the test of equality.

See the section “Binomial Proportion” on page 168 for details.

### Table 3.10 BINOMIAL Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LEVEL=</strong></td>
<td>Specifies the variable level</td>
</tr>
<tr>
<td><strong>P=</strong></td>
<td>Specifies the null proportion</td>
</tr>
<tr>
<td><strong>CORRECT</strong></td>
<td>Requests continuity correction</td>
</tr>
<tr>
<td><strong>Request Confidence Limits</strong></td>
<td></td>
</tr>
<tr>
<td>`CL=AGRESTICOULL</td>
<td>AC`</td>
</tr>
<tr>
<td><code>CL=ALL</code></td>
<td>Requests all confidence limits</td>
</tr>
<tr>
<td>`CL=EXACT</td>
<td>CLOPPERPEARSON`</td>
</tr>
<tr>
<td><code>CL=JEFFREYS</code></td>
<td>Requests Jeffreys confidence limits</td>
</tr>
<tr>
<td><code>CL=WALD</code></td>
<td>Requests Wald confidence limits</td>
</tr>
<tr>
<td>`CL=WILSON</td>
<td>SCORE`</td>
</tr>
<tr>
<td><strong>Request Tests</strong></td>
<td></td>
</tr>
<tr>
<td>`EQUIV</td>
<td>EQUIVALENCE`</td>
</tr>
<tr>
<td>`NONINF</td>
<td>NONINFERIORITY`</td>
</tr>
<tr>
<td>`SUP</td>
<td>SUPERIORITY`</td>
</tr>
<tr>
<td><code>MARGIN=</code></td>
<td>Specifies the test margin</td>
</tr>
<tr>
<td>`VAR=SAMPLE</td>
<td>NULL`</td>
</tr>
</tbody>
</table>

You can specify the following `binomial-options`: 
CL=type | (types)
requests confidence limits for the binomial proportion. You can specify one or more types of confidence limits. When you specify only one type, you can omit the parentheses around the request.

PROC FREQ displays the confidence limits in the “Binomial Confidence Limits” table. The ALPHA= option determines the level of the confidence limits that the CL= binomial-option provides. The default of ALPHA=0.05 produces 95% confidence limits for the binomial proportion. This differs from the test-based confidence limits that are provided with the equivalence, noninferiority, and superiority tests (EQUIV, NONINF, and SUP), which have a confidence coefficient of $100(1 - 2\alpha)\%$ (Schuirmann 1999). See the sections “Equivalence Test” on page 173 and “Noninferiority Test” on page 171 for details.

You can specify the CL= binomial-option with or without requests for binomial tests. The confidence limits that CL= produces do not depend on the tests that you request and do not use the value of the test margin (which you specify by using the MARGIN= binomial-option).

The following types of binomial confidence limits are available:

AGRESTICOULL
AC
requests Agresti-Coull confidence limits for the binomial proportion. See the section “Agresti-Coull Confidence Limits” on page 168 for details.

ALL
requests all available types of confidence limits for the binomial proportion. These include the following: Agresti-Coull, exact (Clopper-Pearson), Jeffreys, Wald, and Wilson (score) confidence limits.

EXACT
CLOPPERPEARSON
requests exact (Clopper-Pearson) confidence limits for the binomial proportion. See the section “Exact (Clopper-Pearson) Confidence Limits” on page 169 for details. If you do not request any binomial confidence limits by specifying the CL= binomial-option, PROC FREQ produces Wald and exact (Clopper-Pearson) confidence limits by default. To request exact tests for the binomial proportion, specify the BINOMIAL option in the EXACT statement.

JEFFREYS
requests Jeffreys confidence limits for the binomial proportion. See the section “Jeffreys Confidence Limits” on page 169 for details.

WALD < (CORRECT) >
requests Wald confidence limits for the binomial proportion. See the section “Wald Confidence Limits” on page 168 for details. If you do not request any binomial confidence limits by specifying the CL= binomial-option, PROC FREQ produces Wald and exact (Clopper-Pearson) confidence limits by default.

If you specify CL=WALD(CORRECT), the Wald confidence limits include a continuity correction. If you specify the CORRECT binomial-option, the Wald confidence limits and the Wald tests include continuity corrections.
WILSON < (CORRECT) >
SCORE <(CORRECT) >
requests Wilson confidence limits for the binomial proportion. These are also known as score confidence limits. See the section “Wilson (Score) Confidence Limits” on page 169 for details.

If you specify CL=WILSON(CORRECT) or the CORRECT binomial-option, the Wilson confidence limits include a continuity correction.

CORRECT
includes a continuity correction in the Wald confidence limits, Wald tests, and Wilson confidence limits.

You can request continuity corrections individually for Wald or Wilson confidence limits by specifying the CL=WALD(CORRECT) or CL=WILSON(CORRECT) binomial-option, respectively.

EQUIV
EQUIVALENCE
requests a test of equivalence for the binomial proportion. See the section “Equivalence Test” on page 173 for details. You can specify the equivalence test margins, the null proportion, and the variance type with the MARGIN=, P=, and VAR= binomial-options, respectively. To request an exact equivalence test, specify the BINOMIAL option in the EXACT statement.

LEVEL=level-number | ‘level-value’
specifies the variable level for the binomial proportion. By default, PROC FREQ computes the proportion of observations for the first variable level that appears in the output. To request a different level, use LEVEL=level-number or LEVEL=’level-value’, where level-number is the variable level’s number or order in the output, and level-value is the formatted value of the variable level. The value of level-number must be a positive integer. You must enclose level-value in single quotes.

MARGIN=value | (lower, upper)
specifies the margin for the noninferiority, superiority, and equivalence tests, which you request with the NONINF, SUP, and EQUIV binomial-options, respectively. If you do not specify this option, PROC FREQ uses a margin of 0.2 by default.

For noninferiority and superiority tests, specify a single value for the MARGIN= option. The MARGIN= value must be a positive number. You can specify value as a number between 0 and 1. Or you can specify value in percentage form as a number between 1 and 100, and PROC FREQ converts that number to a proportion. The procedure treats the value 1 as 1%.

For noninferiority and superiority tests, the test limits must be between 0 and 1. The limits are determined by the null proportion value (which you can specify with the P= binomial-option) and by the margin value. The noninferiority limit equals the null proportion minus the margin. By default, the null proportion equals 0.5 and the margin equals 0.2, which gives a noninferiority limit of 0.3. The superiority limit equals the null proportion plus the margin, which is 0.7 by default.

For an equivalence test, you can specify a single MARGIN= value, or you can specify both lower and upper values. If you specify a single MARGIN= value, it must be a positive number, as described previously. If you specify a single MARGIN= value for an equivalence test, PROC
FREQ uses \(-value\) as the lower margin and \(value\) as the upper margin for the test. If you specify both \(lower\) and \(upper\) values for an equivalence test, you can specify them in proportion form as numbers between \(-1\) or \(1\). Or you can specify them in percentage form as numbers between \(-100\) and \(100\), and PROC FREQ converts the numbers to proportions. The value of \(lower\) must be less than the value of \(upper\).

The equivalence limits must be between 0 and 1. The equivalence limits are determined by the null proportion value (which you can specify with the \(P=\ binomial\)-option) and by the margin values. The lower equivalence limit equals the null proportion plus the lower margin. By default, the null proportion equals 0.5 and the lower margin equals \(-0.2\), which gives a lower equivalence limit of 0.3. The upper equivalence limit equals the null proportion plus the upper margin, which is 0.7 by default.

See the sections “Noninferiority Test” on page 171 and “Equivalence Test” on page 173 for details.

**NONINF**

requests a test of noninferiority for the binomial proportion. See the section “Noninferiority Test” on page 171 for details. You can specify the noninferiority test margin, the null proportion, and the variance type with the \(MARGIN=\), \(P=\), and \(VAR=\ binomial\)-options, respectively. To request an exact noninferiority test, specify the \(BINOMIAL\) option in the \(EXACT\) statement.

**P=**\(value\)

specifies the null hypothesis proportion for the binomial tests. If you omit this option, PROC FREQ uses 0.5 as the null proportion. The null proportion \(value\) must be a positive number. You can specify \(value\) as a number between 0 and 1. Or you can specify \(value\) in percentage form as a number between 1 and 100, and PROC FREQ converts that number to a proportion. The procedure treats the value 1 as 1%.

**SUP**

requests a test of superiority for the binomial proportion. See the section “Superiority Test” on page 172 for details. You can specify the superiority test margin, the null proportion, and the variance type with the \(MARGIN=\), \(P=\), and \(VAR=\ binomial\)-options, respectively. To request an exact superiority test, specify the \(BINOMIAL\) option in the \(EXACT\) statement.

**VAR=**\(SAMPLE\ | NULL\)

specifies the type of variance estimate to use in the Wald tests of noninferiority, superiority, and equivalence. The default is \(VAR=\ SAMPLE\), which estimates the variance from the sample proportion. \(VAR=\ NULL\) uses a test-based variance that is computed from the null hypothesis proportion (which is specified by the \(P=\ binomial\)-option). See the sections “Noninferiority Test” on page 171 and “Equivalence Test” on page 173 for details.

**CELLCHI2**

displays each table cell’s contribution to the Pearson chi-square statistic in the crosstabulation table. The cell chi-square is computed as \((frequency - expected)^2/expected\), where \(frequency\) is the table cell frequency (count) and \(expected\) is the expected cell frequency, which is computed under the null hypothesis that the row and column variables are independent. For more information, see the section “Pearson Chi-Square Test for Two-Way Tables” on page 154. This option has no effect for one-way
tables or for tables that are displayed in list format (which you can request by specifying the LIST option).

**CHISQ < (chisq-options)>**

requests chi-square tests of homogeneity or independence and measures of association that are based on the chi-square statistic. For two-way tables, the chi-square tests include the Pearson chi-square, likelihood ratio chi-square, and Mantel-Haenszel chi-square tests. The chi-square measures include the phi coefficient, contingency coefficient, and Cramér’s $V$. For $2 \times 2$ tables, the CHISQ option also provides Fisher’s exact test and the continuity-adjusted chi-square test. See the section “Chi-Square Tests and Statistics” on page 153 for details.

For one-way tables, the CHISQ option provides the Pearson chi-square goodness-of-fit test. You can also request the likelihood ratio goodness-of-fit test for one-way tables by specifying the LRCHI chisq-option in parentheses after the CHISQ option. By default, the one-way chi-square tests are based on the null hypothesis of equal proportions. Alternatively, you can provide null hypothesis proportions or frequencies by specifying the TESTP= or TESTF= chisq-option, respectively. See the section “Chi-Square Test for One-Way Tables” on page 153 for more information.

To request Fisher’s exact test for tables larger than $2 \times 2$, specify the FISHER option in the EXACT statement. Exact $p$-values are also available for the Pearson, likelihood ratio, and Mantel-Haenszel chi-square tests. See the description of the EXACT statement for more information.
You can specify the following **chisq-options**:

**DF=df**

specifies the degrees of freedom for the chi-square tests. The value of *df* must not be zero. If the value of *df* is positive, PROC FREQ uses *df* as the degrees of freedom for the chi-square tests. If the value of *df* is negative, PROC FREQ uses *df* to adjust the default degrees of freedom for the chi-square tests.

By default for one-way tables, the value of *df* is \((n - 1)\), where *n* is the number of variable levels in the table. By default for two-way tables, the value of *df* is \((r - 1)(c - 1)\), where *r* is the number of rows in the table and *c* is the number of columns. See the sections “Chi-Square Test for One-Way Tables” on page 153 and “Chi-Square Tests and Statistics” on page 153 for more information.

If you specify a negative value of *df*, PROC FREQ adjusts the default degrees of freedom by adding the (negative) value of *df* to the default value to produce the adjusted degrees of freedom. The adjusted degrees of freedom must be positive.

The **DF= chisq-option** specifies or adjusts the degrees of freedom for the following chi-square tests: the Pearson and likelihood ratio goodness-of-fit tests for one-way tables; and the Pearson, likelihood ratio, and Mantel-Haenszel chi-square tests for two-way tables.

**LRCHI**

requests the likelihood ratio goodness-of-fit test for one-way tables. See the section “Likelihood Ratio Chi-Square Test for One-Way Tables” on page 155 for more information.

By default, this test is based on the null hypothesis of equal proportions. You can provide null hypothesis proportions or frequencies by specifying the **TESTP=** or **TESTF=** *chisq-option*, respectively. You can request an exact likelihood ratio goodness-of-fit test by specifying the **LRCHI** option in the **EXACT** statement.

**TESTF=(values) SAS-data-set**

specifies null hypothesis frequencies for the one-way chi-square goodness-of-fit tests. See the section “Chi-Square Test for One-Way Tables” on page 153 for details. You can list the null frequencies as *values* in parentheses after **TESTF=**. Or you can provide the null frequencies in a secondary input data set by specifying **TESTF=SAS-data-set**. The **TESTF=SAS-data-set** cannot be the same data set that you specify in the **DATA=** option. You can specify only one **TESTF=** or **TESTP=** data set in a single invocation of the procedure.

If you list the null frequencies as *values*, you can separate the *values* with blanks or commas. The *values* must be positive numbers. The number of *values* must equal the number of variable levels in the one-way table. The sum of the *values* must equal the total frequency for the one-way table. Order the *values* to match the order in which the corresponding variable levels appear in the one-way frequency table.

If you provide the null frequencies in a secondary input data set (**TESTF=SAS-data-set**), the variable that contains the null frequencies should be named **_TESTF_**, **TestFrequency**, or **Frequency**. The null frequencies must be positive numbers. The number of frequencies must equal the number of levels in the one-way frequency table, and the sum of the frequencies must equal the total frequency for the one-way table. Order the null frequencies in the data set to match the order in which the corresponding variable levels appear in the one-way frequency table.
**TESTP=(values) SAS-data-set**

specifies null hypothesis proportions for the one-way chi-square goodness-of-fit tests. See the section “Chi-Square Test for One-Way Tables” on page 153 for details. You can list the null proportions as *values* in parentheses after TESTP=. Or you can provide the null proportions in a secondary input data set by specifying TESTP=SAS-data-set. The TESTP=SAS-data-set cannot be the same data set that you specify in the DATA= option. You can specify only one TESTF= or TESTP= data set in a single invocation of the procedure.

If you list the null proportions as *values*, you can separate the *values* with blanks or commas. The *values* must be positive numbers. The number of *values* must equal the number of variable levels in the one-way table. Order the *values* to match the order in which the corresponding variable levels appear in the one-way frequency table. You can specify *values* in probability form as numbers between 0 and 1, where the proportions sum to 1. Or you can specify *values* in percentage form as numbers between 0 and 100, where the percentages sum to 100.

If you provide the null proportions in a secondary input data set (TESTP=SAS-data-set), the variable that contains the null proportions should be named _TESTP_, TestPercent, or Percent. The null proportions must be positive numbers. The number of proportions must equal the number of levels in the one-way frequency table. You can provide the proportions in probability form as numbers between 0 and 1, where the proportions sum to 1. Or you can provide the proportions in percentage form as numbers between 0 and 100, where the percentages sum to 100. Order the null proportions in the data set to match the order in which the corresponding variable levels appear in the one-way frequency table.

**WARN=type | (types)**

controls the warning message for the validity of the asymptotic Pearson chi-square test. By default, PROC FREQ displays a warning message when more than 20% of the table cells have expected frequencies that are less than 5. If you specify the NOPRINT option in the PROC FREQ statement, the procedure displays the warning in the log; otherwise, the procedure displays the warning as a footnote in the chi-square table. You can use the WARN= option to suppress the warning and to include a warning indicator in the output data set.

You can specify one or more of the following *types* in the WARN= option. If you specify more than one *type* value, enclose the values in parentheses after WARN=. For example, *warn = (output noprint).*

<table>
<thead>
<tr>
<th>Value of WARN=</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUTPUT</td>
<td>Adds a warning indicator variable to the output data set</td>
</tr>
<tr>
<td>NOLOG</td>
<td>Suppresses the chi-square warning message in the log</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>Suppresses the chi-square warning message in the display</td>
</tr>
<tr>
<td>NONE</td>
<td>Suppresses the chi-square warning message entirely</td>
</tr>
</tbody>
</table>

If you specify the WARN=OUTPUT option, the ODS output data set ChiSq contains a variable named Warning that equals 1 for the Pearson chi-square observation when more than 20% of the table cells have expected frequencies that are less than 5 and equals 0 otherwise. If you specify WARN=OUTPUT and also specify the CHISQ option in the OUTPUT statement, the statistics output data set contains a variable named WARN_PCHI that indicates the warning.

The WARN=NOLOG option has the same effect as the NOWARN option in the TABLES statement.
requests confidence limits for the measures of association, which you can request by specifying the
MEASURES option. For more information, see the sections “Measures of Association” on page 158
and “Confidence Limits” on page 159. You can set the level of the confidence limits by using the
ALPHA= option. The default of ALPHA=0.05 produces 95% confidence limits.

If you omit the MEASURES option, the CL option invokes MEASURES. The CL option is equivalent
to the MEASURES(CL) option.

CMH < (cmh-options) >
requests Cochran-Mantel-Haenszel statistics, which test for association between the row and column
variables after adjusting for the remaining variables in a multiway table. The Cochran-Mantel-Haenszel
statistics include the nonzero correlation statistic, the row mean scores (ANOVA) statistic, and the
general association statistic. In addition, for 2×2 tables, the CMH option provides the adjusted
Mantel-Haenszel and logit estimates of the odds ratio and relative risks, together with their confidence
limits. For stratified 2×2 tables, the CMH option provides the Breslow-Day test for homogeneity of
odds ratios. (To request Tarone’s adjustment for the Breslow-Day test, specify the BDT cmh-option.)

See the section “Cochran-Mantel-Haenszel Statistics” on page 197 for details.

You can use the CMH1 or CMH2 option to control the number of CMH statistics that PROC FREQ
computes.

For stratified 2×2 tables, you can request Zelen’s exact test for equal odds ratios by specifying the
EQOR option in the EXACT statement. See the section “Zelen’s Exact Test for Equal Odds Ratios” on
page 203 for details. You can request exact confidence limits for the common odds ratio by specifying
the COMOR option in the EXACT statement. This option also provides a common odds ratio test. See
the section “Exact Confidence Limits for the Common Odds Ratio” on page 203 for details.

You can specify the following cmh-options in parentheses after the CMH option. These cmh-options,
which apply to stratified 2×2 tables, are also available with the CMH1 or CMH2 option.

BDT
requests Tarone’s adjustment in the Breslow-Day test for homogeneity of odds ratios. See the
section “Breslow-Day Test for Homogeneity of the Odds Ratios” on page 202 for details.

GAILSIMON < (COLUMN=1 | 2)>
GS < (COLUMN=1 | 2)>
requests the Gail-Simon test for qualitative interaction, which applies to stratified 2×2 tables. See the section “Gail-Simon Test for Qualitative Interactions” on page 205 for details.

The COLUMN= option specifies the column of the risk differences to use in computing the Gail-
Simon test. By default, PROC FREQ uses column 1 risk differences. If you specify COLUMN=2,
PROC FREQ uses column 2 risk differences.

The GAILSIMON cmh-option has the same effect as the GAILSIMON option in the TABLES
statement.

MANTELFLEISS
MF
requests the Mantel-Fleiss criterion for the Mantel-Haenszel statistic for stratified 2×2 tables. See the section “Mantel-Fleiss Criterion” on page 200 for details.
CMH1 < (cmh-options)>
requests the Cochran-Mantel-Haenszel correlation statistic. This option does not provide the CMH row mean scores (ANOVA) statistic or the general association statistic, which are provided by the CMH option. For tables larger than $2 \times 2$, the CMH1 option requires less memory than the CMH option, which can require an enormous amount of memory for large tables.

For $2 \times 2$ tables, the CMH1 option also provides the adjusted Mantel-Haenszel and logit estimates of the odds ratio and relative risks, together with their confidence limits. For stratified $2 \times 2$ tables, the CMH1 option provides the Breslow-Day test for homogeneity of odds ratios.

The cmh-options for CMH1 are the same as the cmh-options that are available with the CMH option. See the description of the CMH option for details.

CMH2 < (cmh-options)>
requests the Cochran-Mantel-Haenszel correlation statistic and the row mean scores (ANOVA) statistic. This option does not provide the CMH general association statistic, which is provided by the CMH option. For tables larger than $2 \times 2$, the CMH2 option requires less memory than the CMH option, which can require an enormous amount of memory for large tables.

For $2 \times 2$ tables, the CMH1 option also provides the adjusted Mantel-Haenszel and logit estimates of the odds ratio and relative risks, together with their confidence limits. For stratified $2 \times 2$ tables, the CMH1 option provides the Breslow-Day test for homogeneity of odds ratios.

The cmh-options for CMH2 are the same as the cmh-options that are available with the CMH option. See the description of the CMH option for details.

CONTENTS='string'
specifies the label to use for crosstabulation tables in the contents file, the Results window, and the trace record. For information about output presentation, see the SAS Output Delivery System: User’s Guide.

If you omit the CONTENTS= option, the contents label for crosstabulation tables is “Cross-Tabular Freq Table” by default.

Note that contents labels for all crosstabulation tables that are produced by a single TABLES statement use the same text. To specify different contents labels for different crosstabulation tables, request the tables in separate TABLES statements and use the CONTENTS= option in each TABLES statement.

To remove the crosstabulation table entry from the contents file, you can specify a null label with CONTENTS="."

The CONTENTS= option affects only contents labels for crosstabulation tables. It does not affect contents labels for other PROC FREQ tables.

To specify the contents label for any PROC FREQ table, you can use PROC TEMPLATE to create a customized table template. The CONTENTS_LABEL attribute in the DEFINE TABLE statement of PROC TEMPLATE specifies the contents label for the table. See the chapter “The TEMPLATE Procedure” in the SAS Output Delivery System: User’s Guide for more information.

CROSSLIST < (options)>
displays crosstabulation tables by using an ODS column format instead of the default crosstabulation cell format. In the CROSSLIST table display, the rows correspond to the crosstabulation table cells, and the columns correspond to descriptive statistics such as frequencies and percentages. The CROSSLIST
table displays the same information as the default crosstabulation table (but it uses an ODS column format). For more information about the contents of the CROSSLIST table, see the section “Two-Way and Multiway Tables” on page 216.

You can control the contents of a CROSSLIST table by specifying the same options available for the default crosstabulation table. These include the NOFREQ, NOPERCENT, NOROW, and NOCOL options. You can request additional information in a CROSSLIST table by specifying the CELLCHI2, DEVIATION, EXPECTED, MISSPRINT, and TOTPCT options. You can also display standardized residuals or Pearson residuals in a CROSSLIST table by specifying the CROSSLIST(STDRES) or CROSSLIST(PEARSONRES) option, respectively; these options are not available for the default crosstabulation table. The FORMAT= and CUMCOL options have no effect on CROSSLIST tables.

You cannot specify both the LIST option and the CROSSLIST option in the same TABLES statement.

For CROSSLIST tables, you can use the NOSPARSE option to suppress display of variable levels that have zero frequencies. By default, PROC FREQ displays all levels of the column variable within each level of the row variable, including any levels that have zero frequencies. By default for multiway tables that are displayed as CROSSLIST tables, the procedure displays all levels of the row variable for each stratum of the table, including any row levels with zero frequencies in the stratum.

You can specify the following options:

STDRES
displays the standardized residuals of the table cells in the CROSSLIST table. The standardized residual is the ratio of \( \frac{\text{frequency} - \text{expected}}{\text{standard error}} \), where frequency is the table cell frequency (count) and expected is the expected table cell frequency, which is computed under the null hypothesis that the row and column variables are independent. For more information, see the section “Standardized Residuals” on page 154. You can display the expected values and deviations by specifying the EXPECTED and DEVIATION options, respectively.

PEARSONRES
displays the Pearson residuals of the table cells in the CROSSLIST table. The Pearson residual is the square root of the table cell’s contribution to the Pearson chi-square statistic. The Pearson residual is computed as \( \frac{\text{frequency} - \text{expected}}{\sqrt{\text{expected}}} \), where frequency is the table cell frequency (count) and expected is the expected table cell frequency, which is computed under the null hypothesis that the row and column variables are independent. For more information, see the section “Pearson Chi-Square Test for Two-Way Tables” on page 154. You can display the expected values, deviations, and cell chi-squares by specifying the EXPECTED, DEVIATION, and CELLCHI2 options, respectively.

CUMCOL
displays the cumulative column percentages in the cells of the crosstabulation table. The CUMCOL option does not apply to crosstabulation tables produced with the LIST or CROSSLIST option.

DEVIAITON
displays the deviations of the frequencies from the expected frequencies \( \text{frequency} - \text{expected} \) in the crosstabulation table. The expected frequencies are computed under the null hypothesis that the row and column variables are independent. For more information, see the section “Pearson Chi-Square Test for Two-Way Tables” on page 154. You can display the expected values by specifying the EXPECTED option. This option has no effect for one-way tables or for tables that are displayed in list format (which you can request by specifying the LIST option).
EXPECTED
displays the expected cell frequencies in the crosstabulation table. The expected frequencies are computed under the null hypothesis that the row and column variables are independent. For more information, see the section “Pearson Chi-Square Test for Two-Way Tables” on page 154. This option has no effect for one-way tables or for tables that are displayed in list format (which you can request by specifying the LIST option).

FISHER
requests Fisher’s exact test for tables that are larger than $2 \times 2$. (For $2 \times 2$ tables, the CHISQ option provides Fisher’s exact test.) This test is also known as the Freeman-Halton test. See the sections “Fisher’s Exact Test” on page 157 and “Exact Statistics” on page 206 for more information.

If you omit the CHISQ option in the TABLES statement, the FISHER option invokes CHISQ. You can also request Fisher’s exact test by specifying the FISHER option in the EXACT statement.

NOTE: PROC FREQ computes exact tests by using fast and efficient algorithms that are superior to direct enumeration. Exact tests are appropriate when a data set is small, sparse, skewed, or heavily tied. For some large problems, computation of exact tests might require a substantial amount of time and memory. Consider using asymptotic tests for such problems. Alternatively, when asymptotic methods might not be sufficient for such large problems, consider using Monte Carlo estimation of exact $p$-values. You can request Monte Carlo estimation by specifying the MC computation-option in the EXACT statement. See the section “Computational Resources” on page 208 for more information.

FORMAT= format-name
specifies a format for the following crosstabulation table cell values: frequency, expected frequency, and deviation. PROC FREQ also uses the specified format to display the row and column total frequencies and the overall total frequency in crosstabulation tables.

You can specify any standard SAS numeric format or a numeric format defined with the FORMAT procedure. The format length must not exceed 24. If you omit the FORMAT= option, by default PROC FREQ uses the BEST6. format to display frequencies less than 1E6, and the BEST7. format otherwise.

The FORMAT= option applies only to crosstabulation tables displayed in the default format. It does not apply to crosstabulation tables produced with the LIST or CROSSLIST option.

To change display formats in any FREQ table, you can use PROC TEMPLATE. See the chapter “The TEMPLATE Procedure” in the SAS Output Delivery System: User’s Guide for more information.

GAILSIMON < (COLUMN=1 | 2) >
GS < (COLUMN=1 | 2) >
requests the Gail-Simon test for qualitative interaction, which applies to stratified $2 \times 2$ tables. See the section “Gail-Simon Test for Qualitative Interactions” on page 205 for details.

The COLUMN= option specifies the column of the risk differences to use in computing the Gail-Simon test. By default, PROC FREQ uses column 1 risk differences. If you specify COLUMN=2, PROC FREQ uses column 2 risk differences.

JT
requests the Jonckheere-Terpstra test. See the section “Jonckheere-Terpstra Test” on page 190 for details. To request exact $p$-values for the Jonckheere-Terpstra test, specify the JT option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.
Chapter 3: The FREQ Procedure

LIST

displays two-way and multiway tables by using a list format instead of the default crosstabulation cell format. This option displays an entire multiway table in one table, instead of displaying a separate two-way table for each stratum. For more information, see the section “Two-Way and Multiway Tables” on page 216.

The LIST option is not available when you request tests and statistics; you must use the standard crosstabulation table display or the CROSSLIST display when you request tests and statistics.

MAXLEVELS=n

specifies the maximum number of variable levels to display in one-way frequency tables. The value of n must be a positive integer. PROC FREQ displays the first n variable levels, matching the order in which the levels appear in the one-way frequency table. (The ORDER= option controls the order of the variable levels. By default, ORDER=INTERNAL, which orders the variable levels by unformatted value.)

The MAXLEVELS= option also applies to one-way frequency plots, which you can request by specifying the PLOTS=FREQPLOT option when ODS Graphics is enabled.

If you specify the MISSPRINT option to display missing levels in the frequency table, the MAXLEVELS= option displays the first n nonmissing levels.

The MAXLEVELS= option does not apply to the OUT= output data set, which includes all variable levels. The MAXLEVELS= option does not affect the computation of percentages, statistics, or tests for the one-way table; these values are based on the complete table.

MEASURES < (CL)>

requests measures of association and their asymptotic standard errors. This option provides the following measures: gamma, Kendall’s tau-b, Stuart’s tau-c, Somers’ D(C | R), Somers’ D(R | C), Pearson and Spearman correlation coefficients, lambda (symmetric and asymmetric), and uncertainty coefficients (symmetric and asymmetric). If you specify the CL option in parentheses after the MEASURES option, PROC FREQ provides confidence limits for the measures of association. For more information, see the section “Measures of Association” on page 158.

For 2 x 2 tables, the MEASURES option also provides the odds ratio, column 1 relative risk, column 2 relative risk, and their asymptotic Wald confidence limits. You can request the odds ratio and relative risks separately (without the other measures of association) by specifying the RELRISK option. You can request confidence limits for the odds ratio by specifying the OR(CL=) option.

You can use the TEST statement to request asymptotic tests for the following measures of association: gamma, Kendall’s tau-b, Stuart’s tau-c, Somers’ D(C | R), Somers’ D(R | C), and Pearson and Spearman correlation coefficients. You can use the EXACT statement to request exact confidence limits for the odds ratio, exact unconditional confidence limits for the relative risks, and exact tests for the following measures of association: Kendall’s tau-b, Stuart’s tau-c, Somers’ D(C | R) and D(R | C), and Pearson and Spearman correlation coefficients. For more information, see the descriptions of the TEST and EXACT statements and the section “Exact Statistics” on page 206.

MISSING

treats missing values as a valid nonmissing level for all TABLES variables. The MISSING option displays the missing levels in frequency and crosstabulation tables and includes them in all calculations of percentages, tests, and measures.
By default, if you do not specify the MISSING or MISSPRINT option, an observation is excluded from a table if it has a missing value for any of the variables in the TABLES request. When PROC FREQ excludes observations with missing values, it displays the total frequency of missing observations below the table. See the section “Missing Values” on page 148 for more information.

MISSPRINT

displays missing value frequencies in frequency and crosstabulation tables but does not include the missing value frequencies in any computations of percentages, tests, or measures.

By default, if you do not specify the MISSING or MISSPRINT option, an observation is excluded from a table if it has a missing value for any of the variables in the TABLES request. When PROC FREQ excludes observations with missing values, it displays the total frequency of missing observations below the table. See the section “Missing Values” on page 148 for more information.

NOCOL

suppresses the display of column percentages in crosstabulation table cells.

NOCUM

suppresses the display of cumulative frequencies and percentages in one-way frequency tables. The NOCUM option also suppresses the display of cumulative frequencies and percentages in crosstabulation tables in list format, which you request with the LIST option.

NOFREQ

suppresses the display of cell frequencies in crosstabulation tables. The NOFREQ option also suppresses row total frequencies. This option has no effect for one-way tables or for crosstabulation tables in list format, which you request with the LIST option.

NOPERCENT

suppresses the display of overall percentages in crosstabulation tables. These percentages include the cell percentages of the total (two-way) table frequency, as well as the row and column percentages of the total table frequency. To suppress the display of cell percentages of row or column totals, use the NOROW or NOCOL option, respectively.

For one-way frequency tables and crosstabulation tables in list format, the NOPERCENT option suppresses the display of percentages and cumulative percentages.

NOPRINT

suppresses the display of frequency and crosstabulation tables but displays all requested tests and statistics. To suppress the display of all output, including tests and statistics, use the NOPRINT option in the PROC FREQ statement.

NOROW

suppresses the display of row percentages in crosstabulation table cells.

NOSPARSE

suppresses the display of cells with a zero frequency count in LIST output and omits them from the OUT= data set. The NOSPARSE option applies when you specify the ZEROS option in the WEIGHT statement to include observations with zero weights. By default, the ZEROS option invokes the SPARSE option, which displays table cells with a zero frequency count in the LIST output and includes them in the OUT= data set. See the description of the ZEROS option for more information.
The NOSPARSE option also suppresses the display of variable levels with zero frequency in CROSSLIST tables. By default for CROSSLIST tables, PROC FREQ displays all levels of the column variable within each level of the row variable, including any column variable levels with zero frequency for that row. For multiway tables displayed with the CROSSLIST option, the procedure displays all levels of the row variable for each stratum of the table by default, including any row variable levels with zero frequency for the stratum.

NOWARN

suppresses the log warning message for the validity of the asymptotic Pearson chi-square test. By default, PROC FREQ provides a validity warning for the asymptotic Pearson chi-square test when more than 20 cells have expected frequencies that are less than 5. This warning message appears in the log if you specify the NOPRINT option in the PROC FREQ statement.

The NOWARN option is equivalent to the CHISQ(WARN=NOLOG) option. You can also use the CHISQ(WARN=) option to suppress the warning message in the display and to request a warning variable in the chi-square ODS output data set or in the OUTPUT data set.

OR(CL=type | (types))
ODDSRATIO(CL=type | (types))

requests the odds ratio and confidence limits for $2 \times 2$ tables. You can specify one or more types of confidence limits, which include exact, score, and Wald confidence limits. When you specify only one confidence limit type, you can omit the parentheses around the request.

PROC FREQ displays the confidence limits in the “Confidence Limits for the Odds Ratio” table. Specifying the OR option without the CL= option is equivalent to specifying the RELRISK option, which produces the “Odds Ratio and Relative Risks” table. For more information, see the description of the RELRISK option. When you specify the OR(CL=) option, PROC FREQ does not produce the “Odds Ratio and Relative Risks” table unless you also specify the RELRISK or MEASURES option.

The ALPHA= option determines the confidence level, and the default of ALPHA=0.05 produces 95% confidence limits for the odds ratio.

You can specify the following types:

EXACT

displays exact confidence limits for the odds ratio in the “Confidence Limits for the Odds Ratio” table. You must also request computation of the exact confidence limits by specifying the OR option in the EXACT statement. For more information, see the section “Exact Confidence Limits for the Odds Ratio” on page 187.

SCORE < (CORRECT=NO) >

requests score confidence limits for the odds ratio. For more information, see the section “Score Confidence Limits for the Odds Ratio” on page 187. If you specify CORRECT=NO, PROC FREQ provides the uncorrected form of the confidence limits.

WALD

requests asymptotic Wald confidence limits for the odds ratio. For more information, see the section “Odds Ratio and Relative Risks for 2 x 2 Tables” on page 186.
**OUT=** *SAS-data-set*

names an output data set that contains frequency or crosstabulation table counts and percentages. If more than one table request appears in the TABLES statement, the contents of the OUT= data set correspond to the last table request in the TABLES statement. The OUT= data set variable `COUNT` contains the frequencies and the variable `PERCENT` contains the percentages. See the section “Output Data Sets” on page 211 for details. You can specify the following options to include additional information in the OUT= data set: `OUTCUM`, `OUTEXPECT`, and `OUTPCT`.

**OUTCUM**

includes cumulative frequencies and cumulative percentages in the OUT= data set for one-way tables. The variable `CUM_FREQ` contains the cumulative frequencies, and the variable `CUM_PCT` contains the cumulative percentages. See the section “Output Data Sets” on page 211 for details. The OUTCUM option has no effect for two-way or multiway tables.

**OUTEXPECT**

includes expected cell frequencies in the OUT= data set for crosstabulation tables. The variable `EXPECTED` contains the expected cell frequencies. See the section “Output Data Sets” on page 211 for details. The EXPECTED option has no effect for one-way tables.

**OUTPCT**

includes the following additional variables in the OUT= data set for crosstabulation tables:

- `PCT_COL` percentage of column frequency
- `PCT_ROW` percentage of row frequency
- `PCT_TABL` percentage of stratum (two-way table) frequency, for \( n \)-way tables where \( n > 2 \)

See the section “Output Data Sets” on page 211 for details. The OUTPCT option has no effect for one-way tables.

**PLCORR < (options) >**

requests the polychoric correlation coefficient and its asymptotic standard error. For \( 2 \times 2 \) tables, this statistic is more commonly known as the tetrachoric correlation coefficient, and it is labeled as such in the displayed output. For more information, see the section “Polychoric Correlation” on page 164.

If you also specify the CL or MEASURES(CL) option, PROC FREQ provides confidence limits for the polychoric correlation. If you specify the PLCORR option in the TEST statement, the procedure provides Wald and likelihood ratio tests for the polychoric correlation. The PLCORR option invokes the MEASURES option.

You can specify the following options:

**CONVERGE=** *value*

specifies the convergence criterion for computing the polychoric correlation. The convergence criterion *value* must be a positive number. By default, CONVERGE=0.0001. Iterative computation of the polychoric correlation stops when the convergence measure falls below *value* or when the number of iterations exceeds the MAXITER= *number*, whichever happens first. For parameter values that are less than 0.01, PROC FREQ evaluates convergence by using the absolute difference instead of the relative difference. For more information, see the section “Polychoric Correlation” on page 164.
`MAXITER=number` specifies the maximum number of iterations for computing the polychoric correlation. The value of `number` must be a positive integer. By default, `MAXITER=20`. Iterative computation of the polychoric correlation stops when the number of iterations exceeds the maximum `number` or when the convergence measure falls below the `CONVERGE= value`, whichever happens first. For more information, see the section “Polychoric Correlation” on page 164.

`PLOTS < (global-plot-options) > < =plot-request < (plot-options)> >` controls the plots that are produced through ODS Graphics. `Plot-requests` identify the plots, and `plot-options` control the appearance and content of the plots. You can specify `plot-options` in parentheses after a `plot-request`. A `global-plot-option` applies to all plots for which it is available unless it is altered by a specific `plot-option`. You can specify `global-plot-options` in parentheses after the `PLOTS=` option.

When you specify only one `plot-request`, you can omit the parentheses around the request. For example:

```plaintext
plots=all
plots=freqplot
plots=(freqplot oddsratioplot)
plots(only)=(cumfreqplot deviationplot)
```

ODS Graphics must be enabled before plots can be requested. For example:

```plaintext
ods graphics on;
proc freq;
   tables treatment*response / chisq plots=freqplot;
   weight wt;
run;
ods graphics off;
```


If ODS Graphics is enabled but you do not specify the `PLOTS=` option, PROC FREQ produces all plots that are associated with the analyses that you request, with the exception of the frequency, cumulative frequency, and mosaic plots. To produce a frequency plot or cumulative frequency plot when ODS Graphics is enabled, you must specify the `FREQPLOT` or `CUMFREQPLOT` `plot-request`, respectively, in the `PLOTS=` option, or you must specify the `PLOTS=ALL` option. To produce a mosaic plot when ODS Graphics is enabled, you must specify the `MOSAICPLOT` `plot-request` in the `PLOTS=` option, or you must specify the `PLOTS=ALL` option.

PROC FREQ produces the remaining plots (listed in Table 3.11) by default when you request the corresponding `TABLES` statement options. You can suppress default plots and request specific plots by using the `PLOTS(ONLY)=` option; `PLOTS(ONLY)=(plot-requests)` produces only the plots that are specified as `plot-requests`. You can suppress all plots by specifying the `PLOTS=NONE` option. The `PLOTS` option has no effect when you specify the `NOPRINT` option in the `PROC FREQ` statement.
**Plot Requests**

Table 3.11 lists the available *plot-requests* together with their required TABLES statement options. Descriptions of the *plot-requests* follow the table in alphabetical order.

<table>
<thead>
<tr>
<th>Plot Request</th>
<th>Description</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGREEPLOT</td>
<td>Agreement plot</td>
<td>AGREE (r x r table)</td>
</tr>
<tr>
<td>ALL</td>
<td>All plots</td>
<td>None</td>
</tr>
<tr>
<td>CUMFREQPLOT</td>
<td>Cumulative frequency plot</td>
<td>One-way table request</td>
</tr>
<tr>
<td>DEVIATIONPLOT</td>
<td>Deviation plot</td>
<td>CHISQ (one-way table)</td>
</tr>
<tr>
<td>FREQPLOT</td>
<td>Frequency plot</td>
<td>Any table request</td>
</tr>
<tr>
<td>KAPPAPLOT</td>
<td>Kappa plot</td>
<td>AGREE (h x r x r table)</td>
</tr>
<tr>
<td>MOSAICPLOT</td>
<td>Mosaic plot</td>
<td>Two-way or multiway table request</td>
</tr>
<tr>
<td>NONE</td>
<td>No plots</td>
<td>None</td>
</tr>
<tr>
<td>ODDSRATIOPLOT</td>
<td>Odds ratio plot</td>
<td>MEASURES or RELRISK (h x 2 x 2 table)</td>
</tr>
<tr>
<td>RELRISKPLOT</td>
<td>Relative risk plot</td>
<td>MEASURES or RELRISK (h x 2 x 2 table)</td>
</tr>
<tr>
<td>RISKDIFFPLOT</td>
<td>Risk difference plot</td>
<td>RISKDIFF (h x 2 x 2 table)</td>
</tr>
<tr>
<td>WTKAPPAPLOT</td>
<td>Weighted kappa plot</td>
<td>AGREE (h x r x r table, r &gt; 2)</td>
</tr>
</tbody>
</table>

The following *plot-requests* are available with the PLOTS= option:

**AGREEPLOT < (plot-options) >**

requests an agreement plot (Bangdiwala and Bryan 1987). An agreement plot displays the strength of agreement in a two-way table, where the row and column variables represent two independent ratings of n subjects. For information about agreement plots, see Bangdiwala (1988); Bangdiwala et al. (2008), and Friendly (2000, Section 3.7.2).

To produce an agreement plot, you must also specify the AGREE option in the TABLES statement. Agreement statistics and plots are available for two-way square tables, where the number of rows equals the number of columns.

Table 3.12 lists the *plot-options* that are available for agreement plots. See the subsection “Plot Options” for descriptions of the *plot-options*.

<table>
<thead>
<tr>
<th>Plot Option</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEGEND=</td>
<td>Legend</td>
<td>YES* or NO</td>
</tr>
<tr>
<td>PARTIAL=</td>
<td>Partial agreement</td>
<td>YES* or NO</td>
</tr>
<tr>
<td>SHOWSCALE=</td>
<td>Frequency scale</td>
<td>YES* or NO</td>
</tr>
<tr>
<td>STATS</td>
<td>Statistics</td>
<td>None</td>
</tr>
</tbody>
</table>

*Default
requests all plots that are associated with the specified analyses. Table 3.11 lists the available plot-requests and the corresponding analysis options. If you specify the PLOTS=ALL option, PROC FREQ produces the frequency, cumulative frequency, and mosaic plots that are associated with the tables that you request. (These plots are not produced by default when ODS Graphics is enabled.)

requests a plot of cumulative frequencies. Cumulative frequency plots are available for one-way frequency tables.

To produce a cumulative frequency plot, you must specify the CUMFREQPLOT plot-request in the PLOTS= option, or you must specify the PLOTS=ALL option. PROC FREQ does not produce cumulative frequency plots by default when ODS Graphics is enabled.

Table 3.13 lists the plot-options that are available for cumulative frequency plots. See the subsection “Plot Options” for descriptions of the plot-options.

<table>
<thead>
<tr>
<th>Plot Option</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIENT=</td>
<td>Orientation</td>
<td>VERTICAL* or HORIZONTAL</td>
</tr>
<tr>
<td>SCALE=</td>
<td>Scale</td>
<td>FREQ* or PERCENT</td>
</tr>
<tr>
<td>TYPE=</td>
<td>Type</td>
<td>BARCHART* or DOTPLOT</td>
</tr>
</tbody>
</table>

requests a plot of relative deviations from expected frequencies. Deviation plots are available for chi-square analysis of one-way frequency tables. To produce a deviation plot, you must also specify the CHISQ option in the TABLES statement for a one-way frequency table.

Table 3.14 lists the plot-options that are available for deviation plots. See the subsection “Plot Options” for descriptions of the plot-options.

<table>
<thead>
<tr>
<th>Plot Option</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOSTAT</td>
<td>No statistic</td>
<td>None</td>
</tr>
<tr>
<td>ORIENT=</td>
<td>Orientation</td>
<td>VERTICAL* or HORIZONTAL</td>
</tr>
<tr>
<td>TYPE=</td>
<td>Type</td>
<td>BARCHART* or DOTPLOT</td>
</tr>
</tbody>
</table>

requests a frequency plot. Frequency plots are available for frequency and crosstabulation tables. For multiway crosstabulation tables, PROC FREQ provides a two-way frequency plot for each stratum (two-way table).
To produce a frequency plot, you must specify the FREQPLOT plot-request in the PLOTS= option, or you must specify the PLOTS=ALL option. PROC FREQ does not produce frequency plots by default when ODS Graphics is enabled.

By default, PROC FREQ displays frequency plots as bar charts. You can specify the TYPE=DOTPLOT plot-option to display frequency plots as dot plots. You can plot percentages instead of frequencies by specifying the SCALE=PERCENT plot-option. There are four frequency plot layouts available, which you can request by specifying the TWOWAY= plot-option. See the subsection “Plot Options” for more information.

By default, the primary grouping of graph cells in a two-way layout is by column variable. Row variable levels are then displayed within column variable levels. You can specify the GROUPBY=ROW plot-option to group first by row variable.

Table 3.15 lists the plot-options that are available for frequency plots. See the subsection “Plot Options” for descriptions of the plot-options.

The following plot-options are available for all frequency plots: ORIENT=, SCALE=, and TYPE=. The following plot-options are available for frequency plots of two-way (and multiway) tables: GROUPBY=, NPANELPOS=, and TWOWAY=. The NPANELPOS= plot-option is not available with the TWOWAY=CLUSTER or TWOWAY=STACKED layout, which is always displayed in a single panel.

<table>
<thead>
<tr>
<th>Plot Option</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUPBY=*</td>
<td>Primary group</td>
<td>COLUMN* or ROW</td>
</tr>
<tr>
<td>NPANELPOS=*</td>
<td>Sections per panel</td>
<td>Number (4*)</td>
</tr>
<tr>
<td>ORIENT=</td>
<td>Orientation</td>
<td>VERTICAL* or HORIZONTAL</td>
</tr>
<tr>
<td>SCALE=</td>
<td>Scale</td>
<td>FREQ*, GROUPPERCENT**, LOG, PERCENT, SQRT</td>
</tr>
<tr>
<td>TWOWAY=*</td>
<td>Two-way layout</td>
<td>GROUPVERTICAL* CLUSTER, GROUPHORIZONTAL, or STACKED</td>
</tr>
<tr>
<td>TYPE=</td>
<td>Type</td>
<td>BARCHART* or DOTPLOT</td>
</tr>
</tbody>
</table>

*Default
**For two-way tables

KAPPAPLOT < (plot-options)> requests a plot of kappa statistics with confidence limits. Kappa plots are available for multiway square tables and display the kappa statistic (with confidence limits) for each two-way table (stratum). Kappa plots also display the overall kappa statistic unless you specify the COMMON=NO plot-option. To produce a kappa plot, you must specify the AGREE option in the TABLES statement to compute kappa statistics.

Table 3.16 lists the plot-options that are available for kappa plots. See the subsection “Plot Options” for descriptions of the plot-options.
Table 3.16  Plot Options for KAPPAPLOT and WTKAPPAPLOT

<table>
<thead>
<tr>
<th>Plot Option</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLDISPLAY=</td>
<td>Error bar type</td>
<td>SERIF*, SERIFARROW, LINE, LINEARROW, or BAR</td>
</tr>
<tr>
<td>COMMON=</td>
<td>Overall kappa</td>
<td>YES* or NO</td>
</tr>
<tr>
<td>NPANELPOS=</td>
<td>Statistics per graphic</td>
<td>Number (All*)</td>
</tr>
<tr>
<td>ORDER=</td>
<td>Order of two-way levels</td>
<td>ASCENDING or DESCENDING</td>
</tr>
<tr>
<td>RANGE=</td>
<td>Range to display</td>
<td>Values or CLIP</td>
</tr>
<tr>
<td>STATS</td>
<td>Statistic values</td>
<td>None</td>
</tr>
</tbody>
</table>

*MDefault

MOSAICPLOT < (plot-options)>
requests a mosaic plot. Mosaic plots are available for two-way and multiway crosstabulation tables; for multiway tables, PROC FREQ provides a mosaic plot for each two-way table (stratum).

To produce a mosaic plot, you must specify the MOSAICPLOT plot-request in the PLOTS= option, or you must specify the PLOTS=ALL option. PROC FREQ does not produce mosaic plots by default when ODS Graphics is enabled.

Mosaic plots display tiles that correspond to the crosstabulation table cells. The areas of the tiles are proportional to the frequencies of the table cells. The column variable is displayed on the X axis, and the tile widths are proportional to the relative frequencies of the column variable levels. The row variable is displayed on the Y axis, and the tile heights are proportional to the relative frequencies of the row levels within column levels. For more information, see Friendly (2000).

By default, the colors of the tiles correspond to the row variable levels. If you specify the COLORSTAT= plot-option, the tiles are colored according to the values of the Pearson or standardized residuals.

You can specify the following plot-options:

COLORSTAT < =PEARSONRES | STDRES >
colors the mosaic plot tiles according to the values of residuals. If you specify COLORSTAT=PEARSONRES, the tiles are colored according to the Pearson residuals of the corresponding table cells. For more information, see the section “Pearson Chi-Square Test for Two-Way Tables” on page 154. If you specify COLORSTAT=STDRES, the tiles are colored according to the standardized residuals of the corresponding table cells. For more information, see the section “Standardized Residuals” on page 154. You can display the Pearson or standardized residuals in the CROSSLIST table by specifying the CROSSLIST(PEARSONRES) or CROSSLIST(STDRES) option, respectively.

SQUARE
produces a square mosaic plot, where the height of the Y axis equals the width of the X axis. In a square mosaic plot, the scale of the relative frequencies is the same on both axes. By default, PROC FREQ produces a rectangular mosaic plot.
NONE suppresses all plots.

ODDSRATIO < (plot-options)>
requests a plot of odds ratios with confidence limits. Odds ratio plots are available for multiway 2 × 2 tables and display the odds ratio (with confidence limits) for each 2 × 2 table (stratum). To produce an odds ratio plot, you must also specify the MEASURES, OR(CL=), or RELRISK option in the TABLES statement to compute odds ratios.

By default, the odds ratio plot displays asymptotic Wald confidence limits for the odds ratios (CL=WALD). You can specify the CL=EXACT or CL=SCORE plot-option to display exact or score confidence limits, respectively. To display exact or score confidence limits, you must also request their computation. You can specify the OR option in the EXACT statement to request exact confidence limits for the odds ratio; you can specify the OR(CL=SCORE) option in the TABLES statement to request score confidence limits for the odds ratio.

By default, the odds ratio plot displays the common odds ratio when the procedure computes the common value. You can suppress display of the common odds ratio by specifying the COMMON=NO plot-option. To compute the common odds ratio for a multiway table, you can specify the CMH option in the TABLES statement. To display the common odds ratio when the plot displays exact confidence limits (CL=EXACT), you must request computation of exact confidence limits for the common odds ratio by specifying the COMOR option in the EXACT statement. When you specify the CL=SCORE plot-option, the odds ratio plot does not display the common odds ratio.

Table 3.17 lists the plot-options that are available for odds ratio plots. See the subsection “Plot Options” for descriptions of the plot-options.

<table>
<thead>
<tr>
<th>Plot Option</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL=</td>
<td>Confidence limit type</td>
<td>Type</td>
</tr>
<tr>
<td>CLDISPLAY=</td>
<td>Error bar type</td>
<td>SERIF*, SERIFARROW, LINE, LINEARROW, or BAR</td>
</tr>
<tr>
<td>COMMON=</td>
<td>Common value</td>
<td>YES* or NO</td>
</tr>
<tr>
<td>COLUMN=**</td>
<td>Risk column</td>
<td>1* or 2</td>
</tr>
<tr>
<td>LOGBASE=***</td>
<td>Axis scale</td>
<td>2, E, or 10</td>
</tr>
<tr>
<td>NPANELPOS=</td>
<td>Statistics per graphic</td>
<td>Number (All*)</td>
</tr>
<tr>
<td>ORDER=</td>
<td>Order of two-way levels</td>
<td>ASCENDING or DESCENDING</td>
</tr>
<tr>
<td>RANGE=</td>
<td>Range to display</td>
<td>Values or CLIP</td>
</tr>
<tr>
<td>STATS</td>
<td>Statistic values</td>
<td>None</td>
</tr>
</tbody>
</table>

* Default
** Available for RELRISKPLOT and RISKDIFFPLOT
*** Available for ODDSRATIO and RELRISKPLOT

RELRISKPLOT < (plot-options)>
requests a plot of relative risks with confidence limits. Relative risk plots are available for multiway 2 × 2 tables and display the relative risk (with confidence limits) for each 2 × 2 table.
(stratum). To produce a relative risk plot, you must also specify the MEASURES or RELRISK option in the TABLES statement to compute relative risks.

If you specify the CMH option in the TABLES statement to compute the common relative risk for the multiway table, by default the plot displays the common relative risk. The COMMON=NO plot-option suppresses display of the common relative risk. If you do not specify the CMH option, the common relative risk is not available to be displayed in the plot.

By default, relative risk plots display asymptotic confidence limits. You can specify the CL=EXACT plot-option to display exact confidence limits for the relative risks. You must also request computation of exact confidence limits by specifying the RELRISK option in the EXACT statement. When you specify the CL=EXACT plot-option, the relative risk plot does not display the common relative risk.

Table 3.17 lists the plot-options that are available for relative risk plots. See the subsection “Plot Options” for descriptions of the plot-options.

RISKDIFFPLOT < (plot-options)>
requests a plot of risk (proportion) differences with confidence limits. Risk difference plots are available for multiway $2 \times 2$ tables and display the risk difference (with confidence limits) for each $2 \times 2$ table (stratum). To produce a risk difference plot, you must also specify the RISKDIFF option in the TABLES statement to compute risk differences.

By default, risk difference plots display asymptotic Wald confidence limits. You can specify the CL=EXACT plot-option to display exact confidence limits for the risk differences. You must also request computation of exact confidence limits by specifying the RISKDIFF option in the EXACT statement.

In addition to Wald and exact confidence limits, you can display the following confidence limit types in a risk difference plot: Agresti-Caffo, Hauck-Anderson, Miettinen-Nurminen (score), and Newcombe. See the descriptions of the CL= plot-option and the RISKDIFF option for more information.

By default or when you specify CL=WALD, the risk difference plot displays the common risk difference when PROC FREQ computes the common value. To compute the common risk difference for a multiway table, you can specify the RISKDIFF(COMMON) option in the TABLES statement. When you specify other confidence limit types, the risk difference plot does not display the common risk difference. You can suppress display of the common risk difference by specifying the COMMON=NO plot-option.

Table 3.17 lists the plot-options that are available for risk difference plots. See the subsection “Plot Options” for descriptions of the plot-options.

WTKAPPAPLOT < (plot-options)>
requests a plot of weighted kappa coefficients with confidence limits. Weighted kappa plots are available for multiway square tables and display the weighted kappa coefficient (with confidence limits) for each two-way table (stratum). Weighted kappa plots also display the overall weighted kappa coefficient unless you specify the COMMON=NO plot-option.

To produce a weighted kappa plot, you must specify the AGREE option in the TABLES statement to compute weighted kappa coefficients, and the table dimension must be greater than 1.

Table 3.16 lists the plot-options that are available for weighted kappa plots. See the subsection “Plot Options” for descriptions of the plot-options.
Global Plot Options

A global-plot-option applies to all plots for which the option is available unless it is altered by an individual plot-option. You can specify global-plot-options in parentheses after the PLOTS option. For example:

```plaintext
plots(order=ascending stats)=(riskdiffplot oddsratioplot)
plots(only)=freqplot
```

The following plot-options are available as global-plot-options: CLDISPLAY=, COLUMN=, COMMON=, EXACT, LOGBASE=, NPANELPOS=, ORDER=, ORIENT=, RANGE=, SCALE=, STATS, and TYPE=. See the subsection “Plot Options” for description of these plot-options.

In addition to these plot-options, you can specify the following global-plot-option:

**ONLY**

suppresses the default plots and requests only the plots that are specified as plot-requests.

Plot Options

You can specify the following plot-options in parentheses after a plot-request:

**CL=type**

specifies the type of confidence limits to display. The CL= plot-option is available for the following plots: ODDSRAPIOTPLOT, RELRISKPLOT, and RISKDIFFPLOT.

For odds ratio plots, the available confidence limit types include exact, score, and Wald, which you can request by specifying CL=EXACT, CL=SCORE, and CL=WALD, respectively. The default confidence limit type for odds ratio plots is CL=WALD. When you specify CL=EXACT to display exact confidence limits, you must also request computation of exact confidence limits by specifying the OR option in the EXACT statement. When you specify CL=SCORE, you must request computation of score confidence limits by specifying the OR(CL=SCORE) option in the TABLES statement.

For relative risk plots, the available confidence limit types include exact and Wald, which you can request by specifying CL=EXACT and CL=WALD, respectively. The default is CL=WALD. When you specify CL=EXACT to display exact confidence limits, you must also request computation of exact confidence limits by specifying the RELRISK option in the EXACT statement.

For risk difference plots, the available confidence limit types include the following: CL=AC (Agresti-Caffo), CL=EXACT (exact unconditional), CL=HA (Hauck-Anderson), CL=MN (Miettinen-Nurminen (score)), CL=NEWCOMBE (Newcombe), and CL=WALD (Wald). For more information, see the description of the CL= riskdiff-option and the section “Risk Difference Confidence Limits” on page 176. If you specify CL=EXACT to display exact confidence limits in the plot, you must also request computation of exact confidence limits by specifying the RISKDIFF option in the EXACT statement.
Chapter 3: The FREQ Procedure

CLDISPLAY=SERIF | SERIFARROW | LINE | LINEARROW | BAR < width >
controls the appearance of the confidence limit error bars. This plot-option is available for the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

The default is CLDISPLAY=SERIF, which displays the confidence limits as lines with serifs. CLDISPLAY=LINE displays the confidence limits as plain lines without serifs. The CLDISPLAY=SERIFARROW and CLDISPLAY=LINEARROW plot-options display arrowheads on any error bars that are clipped by the RANGE= plot-option; if an entire error bar is cut from the plot, the plot displays an arrowhead that points toward the statistic.

CLDISPLAY=BAR displays the confidence limits as bars. By default, the width of the bars equals the size of the marker for the estimate. You can control the width of the bars and the size of the marker by specifying the value of width as a percentage of the distance between bars, 0 < width ≤ 1. The bar might disappear when the value of width is very small.

COLUMN=1 | 2
specifies the table column to use to compute the risks (proportion) for the relative risk plot (RELRISKPLOT) and the risk difference plot (RISKDIFFPLOT). If you specify COLUMN=1, the plot displays the column 1 relative risks or the column 1 risk differences. Similarly, if you specify COLUMN=2, the plot displays the column 2 relative risks or risk differences.

For relative risk plots, the default is COLUMN=1. For risk difference plots, the default is COLUMN=1 if you request computation of both column 1 and column 2 risk differences by specifying the RISKDIFF option. If you request computation of only the column 1 (or column 2) risk differences by specifying the RISKDIFF(COLUMN=1) (or RISKDIFF(COLUMN=2)) option, by default the risk difference plot displays the risk differences for the column that you specify.

COMMON=YES | NO
controls the display of the common (overall) statistic in plots that display stratum (two-way table) statistics for a multiway table. The COMMON= plot-option is available for the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

By default, COMMON=YES and plots display the common value when PROC FREQ computes it. When you specify the AGREE option in the TABLES statement for a multiway table, PROC FREQ computes the common simple and weighted kappa coefficients, which can be displayed in the simple and weighted kappa plots, respectively. When you specify the MEASURES or RELRISK option in the TABLES statement, PROC FREQ computes only stratum-level odds ratios and relative risks; to compute common odds ratios and relative risks, you must also specify the CMH option in the TABLES statement. To compute the common risk difference, you must specify the RISKDIFF(COMMON) option. For more information, see the descriptions of the plot-requests.

EXACT
requests display of exact confidence limits instead of asymptotic confidence limits. The EXACT plot-option is available for the odds ratio plot (ODDSRATIOPLOT), relative risk plot (RELRISKPLOT), and risk difference plot (RISKDIFFPLOT). The EXACT plot-option is equivalent to the CL=EXACT plot-option.

When you specify the EXACT plot-option, you must also request computation of exact confidence limits by specifying the appropriate statistic-option in the EXACT statement.
GROUPBY=COLUMN | ROW
specifies the primary grouping for two-way frequency plots, which you can request by specifying
the FREQPLOT plot=request. The default is GROUPBY=COLUMN, which groups graph cells
first by column variable and displays row variable levels within column variable levels. You can
specify GROUPBY=ROW to group first by row variable. In two-way and multiway table requests,
the column variable is the last variable specified and forms the columns of the crosstabulation
table. The row variable is the next-to-last variable specified and forms the rows of the table.

By default for a bar chart that is displayed in the TWOWAY=STACKED layout, bars correspond
to the column variable levels, and row levels are displayed (stacked) within each column bar.
By default for a bar chart that is displayed in the TWOWAY=CLUSTER layout, bars are first
grouped by column variable levels, and row levels are displayed as adjacent bars within each
column-level group. You can reverse the default row and column variable grouping by specifying
GROUPBY=ROW.

LOGBASE=2 | E | 10
applies to the odds ratio plot (ODDSRATIO PLOT) and the relative risk plot (RELRISK PLOT).
This plot-option displays the odds ratio or relative risk axis on the log scale that you specify.

LEGEND=YES | NO
applies to the agreement plot (AGREE PLOT). LEGEND=NO suppresses the legend that identifies
the areas of exact and partial agreement. The default is LEGEND=YES.

NOSTAT
applies to the deviation plot (DEVIATION PLOT). NOSTAT suppresses the chi-square p-value
that deviation plot displays by default.

NPANELPOS=n
divides the plot into multiple panels that display at most |n| statistics or sections.
If n is positive, the number of statistics or sections per panel is balanced; if n is negative, the
number of statistics per panel is not balanced. For example, suppose you want to display 21 odds
ratios. NPANELPOS=20 displays two panels, the first with 11 odds ratios and the second with
10 odds ratios; NPANELPOS=–20 displays 20 odds ratios in the first panel but only 1 odds ratio
in the second panel. This plot-option is available for all plots except mosaic plots and one-way
weighted frequency plots.

For two-way frequency plots (FREQ PLOT), NPANELPOS=n requests that panels display at most
|n| sections, where sections correspond to row or column variable levels, depending on the type
of plot and the grouping. By default, n=4 and each panel includes at most four sections. This
plot-option applies to two-way plots that are displayed in the TWOWAY=GROUP VERTICAL or
TWOWAY=GROUP HORIZONTAL layout. The NPANELPOS= plot-option does not apply to the
TWOWAY=CLUSTER and TWOWAY=STACKED layouts, which are always displayed in a
single panel.

For plots that display statistics along with confidence limits, NPANELPOS=n requests that
panels display at most |n| statistics. By default, n=0 and all statistics are displayed in a single
panel. This plot-option applies to the following plots: KAPPAPLOT, OR PLOT, RELRISK PLOT,
RISKDIFF PLOT, and WTKAPPAPLOT.
ORDER=ASCENDING | DESCENDING

displays the two-way table (strata) statistics in order of the statistic value. If you specify ORDER=ASCENDING or ORDER=DESCENDING, the plot displays the statistics in ascending or descending order, respectively. By default, the order of the statistics in the plot matches the order that the two-way table strata appear in the multiway table display.

This plot-option applies to the following plots: KAPPAPLOT, ODDSRATIO PLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

ORIENT=HORIZONTAL | VERTICAL

controls the orientation of the plot. The ORIENT= plot-option applies to the following plots: CUMFREQPLOT, DEVIATIONPLOT, and FREQPLOT.

ORIENT=HORIZONTAL places the variable levels on the Y axis and the frequencies, percentages, or statistic values on the X axis. ORIENT=VERTICAL places the variable levels on the X axis. The default orientation is ORIENT=VERTICAL for bar charts (TYPE=BARCHART) and ORIENT=HORI ZONTAL for dot plots (TYPE=DOTPLOT).

PARTIAL=YES | NO

controls the display of partial agreement in the agreement plot (AGREEPLOT). PARTIAL=NO suppresses the display of partial agreement. When you specify PARTIAL=NO, the agreement plot displays only exact agreement. Exact agreement includes the diagonal cells of the square table, where the row and column variable levels are the same. Partial agreement includes the adjacent off-diagonal table cells, where the row and column values are within one level of exact agreement. The default is PARTIAL=YES.

RANGE=( < min > , max > ) CLIP

specifies the range of values to display. If you specify RANGE=CLIP, the confidence limits are clipped and the display range is determined by the minimum and maximum values of the statistics. By default, the display range includes all confidence limits.

This plot-option is available for the following plots: KAPPAPLOT, ODDSRATIO PLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

SCALE=FREQ | GROUPPERCENT | LOG | PERCENT | SQRT

specifies the scale of the frequencies to display. This plot-option is available for frequency plots (FREQPLOT) and cumulative frequency plots (CUMFREQPLOT).

The default is SCALE=FREQ, which displays unscaled frequencies. SCALE=PERCENT displays percentages (relative frequencies) of the total frequency. SCALE=LOG displays log (base 10) frequencies. SCALE=SQRT displays square roots of the frequencies, producing a plot known as a rootogram.

SCALE=GROUPPERCENT is available for two-way frequency plots. This option displays the row or column percentages instead of the overall percentages (of the table frequency). By default (or when you specify the GROUPBY=COLUMN plot-option), SCALE=GROUPPERCENT displays the column percentages. If you specify the GROUPBY=ROW plot-option, the primary grouping of graph cells is by row variable level and the plot displays row percentages. For more information, see the description of the GROUPBY= plot-option.
SHOWSCALE=YES | NO
controls the display of the cumulative frequency scale on the right side of the agreement plot (AGREEPLOT). SHOWSCALE=NO suppresses the display of the scale. The default is SHOWSCALE=YES.

STATS
displays statistic values in the plot. For the following plots, the STATS plot-option displays the statistics and their confidence limits on the right side of the plot: KAPPAPLOT, ODDSRATIOPILOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

For the agreement plot (AGREEPLOT), STATS displays the values of the kappa statistic, the weighted kappa statistic, and the $B_n$ measure (Bangdiwala and Bryan 1987).

If you do not request the STATS plot-option, these plots do not display the statistic values.

TWOWAY=CLUSTER | GROUPHORIZONTAL | GROUPVERTICAL | STACKED
specifies the layout for two-way frequency plots.

All TWOWAY= layouts are available for bar charts (TYPE=BARCHART). All TWOWAY= layouts except TWOWAY=CLUSTER are available for dot plots (TYPE=DOTPLOT). The ORIENT= and GROUPBY= plot-options are available for all TWOWAY= layouts.

The default two-way layout is TWOWAY=GROUPVERTICAL, which produces a grouped plot that has a vertical common baseline. By default for bar charts (TYPE=BARCHART, ORIENT=VERTICAL), the X axis displays column variable levels, and the Y axis displays frequencies. The plot includes a vertical (Y-axis) block for each row variable level. The relative positions of the graph cells in this plot layout are the same as the relative positions of the table cells in the crosstabulation table. You can reverse the default row and column grouping by specifying the GROUPBY=ROW plot-option.

The TWOWAY=GROUPHORIZONTAL layout produces a grouped plot that has a horizontal common baseline. By default (GROUPBY=COLUMN), the plot displays a block on the X axis for each column variable level. Within each column-level block, the plot displays row variable levels.

The TWOWAY=STACKED layout produces stacked displays of frequencies. By default (GROUPBY=COLUMN) in a stacked bar chart, the bars correspond to column variable levels, and row levels are stacked within each column level. By default in a stacked dot plot, the dotted lines correspond to column levels, and cell frequencies are plotted as data dots on the corresponding column line. The dot color identifies the row level.

The TWOWAY=CLUSTER layout, which is available only for bar charts, displays groups of adjacent bars. By default, the primary grouping is by column variable level, and row levels are displayed within each column level.

You can reverse the default row and column grouping in any layout by specifying the GROUPBY=ROW plot-option. The default is GROUPBY=COLUMN, which groups first by column variable.

TYPE=BARCHART | DOTPLOT
specifies the plot type (format) of the frequency (FREQPLOT), cumulative frequency (CUMFREQPLOT), and deviation plots (DEVIAITIONPLOT). TYPE=BARCHART produces a bar chart and TYPE=DOTPLOT produces a dot plot. The default is TYPE=BARCHART.
PRINTKWTS

Displays the agreement weights that PROC FREQ uses to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. By default, PROC FREQ uses the Cicchetti-Allison form of agreement weights. If you specify the AGREE(WT=FC) option, the procedure uses the Fleiss-Cohen form of agreement weights. For more information, see the section “Weighted Kappa Coefficient” on page 194.

This option has no effect unless you also specify the AGREE option to compute the weighted kappa coefficient. The PRINTKWTS option is equivalent to the AGREE(PRINTKWTS) option.

RELRISK

Or

Requests relative risk measures and their asymptotic Wald confidence limits for \(2 \times 2\) tables. These measures include the odds ratio, the column 1 relative risk, and the column 2 relative risk. PROC FREQ displays the statistics and confidence limits in the “Odds Ratio and Relative Risks” table. For more information, see the section “Odds Ratio and Relative Risks for \(2 \times 2\) Tables” on page 186.

You can request exact confidence limits for the odds ratio by specifying the OR option in the EXACT statement. You can request exact unconditional confidence limits for the relative risks by specifying the RELRISK option in the EXACT statement. For more information, see the sections “Exact Confidence Limits for the Odds Ratio” on page 187 and “Exact Unconditional Confidence Limits for the Relative Risk” on page 188.

You can request score confidence limits for the odds ratio by specifying the OR(CL=) option. This option produces the “Confidence Limits for the Odds Ratio” table, which can also display Wald and exact confidence limits. For more information, see the section “Score Confidence Limits for the Odds Ratio” on page 187.

You can also obtain the relative risk measures by specifying the MEASURES option, which produces other measures of association in addition to the relative risks.

RISKDIFF \(<\ (\text{riskdiff-options})\ >\)

Requests risks (binomial proportions) and risk differences for \(2 \times 2\) tables. When you specify the RISKDIFF option, PROC FREQ provides the row 1 risk, row 2 risk, total (overall) risk, and risk difference (row 1 – row 2), together with their asymptotic standard errors and Wald confidence limits. PROC FREQ also provides exact (Clopper-Pearson) confidence limits for the row 1, row 2, and total risks by default. You can request exact unconditional confidence limits for the risk difference by specifying the RISKDIFF option in the EXACT statement. For more information, see the section “Risks and Risk Differences” on page 174. PROC FREQ displays these results in the column 1 and column 2 “Risk Estimates” tables.

You can specify riskdiff-options in parentheses after the RISKDIFF option to request tests and additional confidence limits for the risk difference, as well as estimates of the common risk difference for multiway \(2 \times 2\) tables. Table 3.18 summarizes the riskdiff-options.

The CL= riskdiff-option requests confidence limits for the risk difference. Available confidence limit types include Agresti-Caffo, exact unconditional, Hauck-Anderson, Miettinen-Nurminen (score), Newcombe, and Wald. Continuity-corrected Newcombe and Wald confidence limits are also available. You can request more than one type of confidence limits in the same analysis. PROC FREQ displays the confidence limits in the “Confidence Limits for the Proportion (Risk) Difference” table.
The CL=EXACT riskdiff-option displays exact unconditional confidence limits in the “Proportion (Risk) Difference Confidence Limits” table. When you use CL=EXACT, you must also request computation of the exact confidence limits by specifying the RISKDIFF option in the EXACT statement.

The EQUIV, NONINF, and SUP riskdiff-options request tests of equivalence, noninferiority, and superiority, respectively, for the risk difference. Available test methods include Farrington-Manning (score), Hauck-Anderson, and Newcombe (hybrid-score), in addition to the Wald test.

As part of the noninferiority, superiority, and equivalence analyses, PROC FREQ provides null-based equivalence limits that have a confidence coefficient of 100(1 − 2α)% (Schuirmann 1999). The ALPHA= option determines the confidence level, and the default of ALPHA=0.05 produces 90% equivalence limits. See the sections “Noninferiority Tests” on page 180 and “Equivalence Tests” on page 182 for details.

### Table 3.18 RISKDIFF (Proportion Difference) Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLUMN=1</td>
<td>2</td>
</tr>
<tr>
<td>COMMON</td>
<td>Requests common risk difference</td>
</tr>
<tr>
<td>CORRECT</td>
<td>Requests continuity correction</td>
</tr>
<tr>
<td>NORISKS</td>
<td>Suppresses default risk tables</td>
</tr>
</tbody>
</table>

#### Request Confidence Limits

- CL=AC: Requests Agresti-Caffo confidence limits
- CL=EXACT: Displays exact confidence limits
- CL=HA: Requests Hauck-Anderson confidence limits
- CL=MN | SCORE: Requests Miettinen-Nurminen confidence limits
- CL=NEWCOMBE: Requests Newcombe confidence limits
- CL=WALD: Requests Wald confidence limits

#### Request Tests

- EQUAL: Requests an equality test
- EQUIV | EQUIVALENCE: Requests an equivalence test
- NONINF | NONINFERIORITY: Requests a noninferiority test
- SUP | SUPERIORITY: Requests a superiority test
- MARGIN=: Specifies the test margin
- METHOD=: Specifies the test method
- VAR=SAMPLE | NULL: Specifies the test variance

You can specify the following riskdiff-options in parentheses after the RISKDIFF option:

- **CL=type | (types)** requests confidence limits for the risk difference. You can specify one or more types of confidence limits. When you specify only one type, you can omit the parentheses around the request. PROC FREQ displays the confidence limits in the “Proportion (Risk) Difference Confidence Limits” table.

The ALPHA= option determines the confidence level, and the default of ALPHA=0.05 produces 95% confidence limits for the risk difference. This differs from the test-based confidence limits that are provided with the equivalence, noninferiority, and superiority tests (EQUIV, NONINF, and SUP), which have a confidence coefficient of 100(1 − 2α)% (Schuirmann 1999).
You can specify CL= with or without requests for risk difference tests. The confidence limits produced by CL= do not depend on the tests that you request and do not use the value of the test margin (which you specify by using MARGIN= riskdiff-option).

You can control the risk column for the confidence limits by specifying the COLUMN= riskdiff-option. If you do not specify COLUMN=, PROC FREQ provides confidence limits for the column 1 risk difference by default.

The following types of confidence limits are available:

**AGRESTICAFFO**

requests Agresti-Caffo confidence limits for the risk difference. See the subsection “Agresti-Caffo Confidence Limits” in the section “Risk Difference Confidence Limits” on page 176 for details.

**EXACT**

displays exact unconditional confidence limits for the risk difference in the “Proportion (Risk) Difference Confidence Limits” table. You must also request computation of the exact confidence limits by specifying the RISKDIFF option in the EXACT statement.

PROC FREQ computes the confidence limits by inverting two separate one-sided exact tests (tail method), where the tests are based on the unstandardized risk difference by default. If you specify the RISKDIFF(METHOD=SCORE) option in the EXACT statement, the tests are based on the score statistic. See the RISKDIFF option in the EXACT statement and the section “Exact Unconditional Confidence Limits for the Risk Difference” on page 183 for more information.

By default, PROC FREQ also displays these exact confidence limits in the “Risk Estimates” table. You can suppress this table by specifying the NORISKS riskdiff-option.

**HA**

requests Hauck-Anderson confidence limits for the risk difference. See the subsection “Hauck-Anderson Confidence Limits” in the section “Risk Difference Confidence Limits” on page 176 for details.

**MN < (CORRECT=NO | MEE) >**

**SCORE < (CORRECT=NO | MEE) >**

requests Miettinen-Nurminen (score) confidence limits for the risk difference. For more information, see the subsection “Miettinen-Nurminen (Score) Confidence Limits” in the section “Risk Difference Confidence Limits” on page 176. By default, the Miettinen-Nurminen confidence limits include a bias correction factor (Miettinen and Nurminen 1985; Newcombe and Nurminen 2011). If you specify CL=MN(CORRECT=NO), PROC FREQ provides the uncorrected form of the confidence limits (Mee 1984).

**NEWCOMBE < (CORRECT) >**

requests Newcombe hybrid-score confidence limits for the risk difference. If you specify CL=NEWCOMBE(CORRECT) or the CORRECT riskdiff-option, the Newcombe confidence limits include a continuity correction. See the subsection “Newcombe Confidence Limits” in the section “Risk Difference Confidence Limits” on page 176 for details.
WALD < (CORRECT) >
requests Wald confidence limits for the risk difference. If you specify CL=WALD(CORRECT) or the CORRECT riskdiff-option, the Wald confidence limits include a continuity correction. See the subsection “Wald Confidence Limits” in the section “Risk Difference Confidence Limits” on page 176 for details.

COLUMN=1 | 2 | BOTH
specifies the table column for which to compute the risk difference tests (EQUAL, EQUIV, NONINF, and SUP) and the risk difference confidence limits (which are requested by the CL= riskdiff-option).

If you do not specify COLUMN=, PROC FREQ provides the risk difference tests and confidence limits for column 1 by default. The COLUMN= option has no effect on the “Risk Estimates” table, which is produced for both column 1 and column 2. You can suppress the “Risk Estimates” table by specifying the NORISKS riskdiff-option.

COMMON
requests estimates of the common (overall) risk difference for multiway $2 \times 2$ tables. PROC FREQ produces Mantel-Haenszel and summary score estimates for the common risk difference, together with their confidence limits. For more information, see the section “Common Risk Difference” on page 185.

If you do not specify the COLUMN= riskdiff-option, PROC FREQ provides the common risk difference for column 1 by default. If you specify COLUMN=2, PROC FREQ provides the common risk difference for column 2. COLUMN=BOTH does not apply to the common risk difference.

CORRECT
includes a continuity correction in the Wald confidence limits, Wald tests, and Newcombe confidence limits. See the section “Risks and Risk Differences” on page 174 for details.

EQUAL
requests a test of the null hypothesis that the risk difference equals zero. PROC FREQ provides an asymptotic Wald test of equality. If you specify the CORRECT riskdiff-option, the Wald test includes a continuity correction. If you specify the VAR=NULL riskdiff-option, the test uses the null (test-based) variance instead of the sample variance. See the section “Equality Test” on page 179 for details.

EQUIV
requests a test of equivalence for the risk difference. See the section “Equivalence Tests” on page 182 for details. You can specify the equivalence test margins with the MARGIN= riskdiff-option and the test method with the METHOD= riskdiff-option. PROC FREQ uses METHOD=WALD by default.

MARGIN=value | (lower, upper)
specifies the margin for the noninferiority, superiority, and equivalence tests, which you request with the NONINF, SUP, and EQUIV riskdiff-options, respectively. If you do not specify MARGIN=, PROC FREQ uses a margin of 0.2 by default.
For noninferiority and superiority tests, specify a single \textit{value} for MARGIN=. The MARGIN= \textit{value} must be a positive number. You can specify \textit{value} as a number between 0 and 1. Or you can specify \textit{value} in percentage form as a number between 1 and 100, and PROC FREQ converts that number to a proportion. The procedure treats the value 1 as 1%.

For an equivalence test, you can specify a single MARGIN= \textit{value}, or you can specify both \textit{lower} and \textit{upper} values. If you specify a single MARGIN= \textit{value}, it must be a positive number, as described previously. If you specify a single MARGIN= \textit{value} for an equivalence test, PROC FREQ uses $-\textit{value}$ as the lower margin and \textit{value} as the upper margin for the test. If you specify both \textit{lower} and \textit{upper} values for an equivalence test, you can specify them in proportion form as numbers between –1 and 1. Or you can specify them in percentage form as numbers between –100 and 100, and PROC FREQ converts the numbers to proportions. The value of \textit{lower} must be less than the value of \textit{upper}.

\textbf{METHOD=method}

specifies the method for the noninferiority, superiority, and equivalence analyses, which you request with the NONINF, SUP, and EQUIV \textit{riskdiff-options}, respectively. If you do not specify the METHOD= \textit{riskdiff-option}, PROC FREQ uses METHOD=WALD by default.

The following \textit{methods} are available:

\textbf{FM}

\textbf{SCORE}

requests Farrington-Manning (score) tests and equivalence limits for the equivalence, noninferiority, and superiority analyses. See the subsection “Farrington-Manning (Score) Test” in the section “Noninferiority Tests” on page 180 for details.

\textbf{HA}

requests Hauck-Anderson tests and confidence limits for the equivalence, noninferiority, and superiority analyses. See the subsection “Hauck-Anderson Test” in the section “Noninferiority Tests” on page 180 for details.

\textbf{NEWCOMBE}

requests Newcombe (hybrid-score) confidence limits for the equivalence, noninferiority, and superiority analyses. If you specify the CORRECT \textit{riskdiff-option}, the Newcombe confidence limits include a continuity correction. See the subsection “Newcombe Noninferiority Analysis” in the section “Noninferiority Tests” on page 180 for details.

\textbf{WALD}

requests Wald tests and confidence limits for the equivalence, noninferiority, and superiority analyses. If you specify the CORRECT \textit{riskdiff-option}, the Wald confidence limits include a continuity correction. If you specify the VAR=NULL \textit{riskdiff-option}, the tests and confidence limits use the null (test-based) variance instead of the sample variance. See the subsection “Wald Test” in the section “Noninferiority Tests” on page 180 for details.

\textbf{NONINF}

\textbf{NONINFERIORITY}

requests a test of noninferiority for the risk difference. See the section “Noninferiority Tests” on page 180 for details. You can specify the test margin with the MARGIN= \textit{riskdiff-option} and the test method with the METHOD= \textit{riskdiff-option}. PROC FREQ uses METHOD=WALD by default.
NORISKS
suppresses display of the “Risk Estimates” tables, which the RISKDIFF option produces by default for column 1 and column 2. The “Risk Estimates” tables contain the risks and risk differences, together with their asymptotic standard errors, Wald confidence limits, and exact confidence limits.

SUP
SUPERIORITY
requests a test of superiority for the binomial proportion. See the section “Superiority Test” on page 182 for details. You can specify the test margin with the MARGIN= riskdiff-option and the test method with the METHOD= riskdiff-option. PROC FREQ uses METHOD=WALD by default.

VAR=SAMPLE | NULL
specifies the type of variance estimate to use in the Wald tests of noninferiority, superiority, equivalence, and equality. The default is VAR=SAMPLE, which uses the sample variance. VAR=NULL uses a test-based variance that is computed by using the null hypothesis risk difference value. See the sections “Equality Test” on page 179 and “Noninferiority Tests” on page 180 for details.

SCORES=type
specifies the type of row and column scores that PROC FREQ uses to compute the following statistics: Mantel-Haenszel chi-square, Pearson correlation, Cochran-Armitage test for trend, weighted kappa coefficient, and Cochran-Mantel-Haenszel statistics. The value of type can be one of the following:

- MODRIDIT
- RANK
- RIDIT
- TABLE

See the section “Scores” on page 152 for descriptions of these score types.

If you do not specify the SCORES= option, PROC FREQ uses SCORES=TABLE by default. For character variables, the row and column TABLE scores are the row and column numbers. That is, the TABLE score is 1 for row 1, 2 for row 2, and so on. For numeric variables, the row and column TABLE scores equal the variable values. See the section “Scores” on page 152 for details. Using MODRIDIT, RANK, or RIDIT scores yields nonparametric analyses.

You can use the SCOROUT option to display the row and column scores.

SCOROUT
displays the row and column scores that PROC FREQ uses to compute score-based tests and statistics. You can specify the score type with the SCORES= option. See the section “Scores” on page 152 for details.

The scores are computed and displayed only when PROC FREQ computes statistics for two-way tables. You can use ODS to store the scores in an output data set. See the section “ODS Table Names” on page 222 for more information.
SPARSE reports all possible combinations of the variable values for an \( n \)-way table when \( n > 1 \), even if a combination does not occur in the data. The SPARSE option applies only to crosstabulation tables displayed in LIST format and to the OUT= output data set. If you do not use the LIST or OUT= option, the SPARSE option has no effect.

When you specify the SPARSE and LIST options, PROC FREQ displays all combinations of variable values in the table listing, including those with a frequency count of zero. By default, without the SPARSE option, PROC FREQ does not display zero-frequency levels in LIST output. When you use the SPARSE and OUT= options, PROC FREQ includes empty crosstabulation table cells in the output data set. By default, PROC FREQ does not include zero-frequency table cells in the output data set.

See the section “Missing Values” on page 148 for more information.

TOTPCT displays the percentage of the total multiway table frequency in crosstabulation tables for \( n \)-way tables, where \( n > 2 \). By default, PROC FREQ displays the percentage of the individual two-way table frequency but does not display the percentage of the total frequency for multiway crosstabulation tables. See the section “Two-Way and Multiway Tables” on page 216 for more information.

The percentage of total multiway table frequency is displayed by default when you specify the LIST option. It is also provided by default in the PERCENT variable in the OUT= output data set.

TREND requests the Cochran-Armitage test for trend. The table must be \( 2 \times C \) or \( R \times 2 \) to compute the trend test. See the section “Cochran-Armitage Test for Trend” on page 189 for details. To request exact \( p \)-values for the trend test, specify the TREND option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

TEST Statement

\[
\text{TEST test-options ;}
\]

The TEST statement requests asymptotic tests for measures of association and measures of agreement. The test-options identify which tests to compute. Table 3.19 lists the available test-options, together with their corresponding TABLES statement options. Descriptions of the test-options follow the table in alphabetical order.

For each measure of association or agreement that you request in the TEST statement, PROC FREQ provides an asymptotic test that the measure equals zero. The procedure displays the asymptotic standard error under the null hypothesis, the test statistic, and the one-sided and two-sided \( p \)-values. Additionally, PROC FREQ reports the confidence limits for the measure. The ALPHA= option in the TABLES statement determines the confidence level, which by default equals 0.05 and provides 95% confidence limits. See the sections “Asymptotic Tests” on page 159 and “Confidence Limits” on page 159 for details. Also see the sections “Measures of Association” on page 158 and “Tests and Measures of Agreement” on page 192 for information about the individual measures.

You can also request exact tests for selected measures of association and agreement by using the EXACT statement. See the section “Exact Statistics” on page 206 for more information.
**Using the TEST Statement with the TABLES Statement**

You must use a TABLES statement with the TEST statement. If you use only one TABLES statement, you do not need to specify the same options in both the TABLES and TEST statements; when you specify an option in the TEST statement, PROC FREQ automatically invokes the corresponding TABLES statement option. However, when you use the TEST statement with multiple TABLES statements, you must specify options in the TABLES statements to request statistics. PROC FREQ then provides asymptotic tests for those statistics that you also specify in the TEST statement.

### Table 3.19  TEST Statement Options

<table>
<thead>
<tr>
<th>Test Option</th>
<th>Asymptotic Tests</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGREE</td>
<td>Simple and weighted kappa coefficients</td>
<td>AGREE</td>
</tr>
<tr>
<td>GAMMA</td>
<td>Gamma</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>KAPPA</td>
<td>Simple kappa coefficient</td>
<td>AGREE</td>
</tr>
<tr>
<td>KENTB</td>
<td>Kendall’s tau-b</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>MEASURES</td>
<td>Gamma, Kendall’s tau-b, Stuart’s tau-c, Somers’ $D(C</td>
<td>R)$, Somers’ $D(R</td>
</tr>
<tr>
<td>PCORR</td>
<td>Pearson correlation coefficient</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>PLCORR</td>
<td>Polychoric correlation</td>
<td>PLCORR</td>
</tr>
<tr>
<td>SCORR</td>
<td>Spearman correlation coefficient</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>SMDCR</td>
<td>Somers’ $D(C</td>
<td>R)$</td>
</tr>
<tr>
<td>SMDRC</td>
<td>Somers’ $D(R</td>
<td>C)$</td>
</tr>
<tr>
<td>STUTC</td>
<td>Stuart’s tau-c</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>WTKAP</td>
<td>Weighted kappa coefficient</td>
<td>AGREE</td>
</tr>
</tbody>
</table>

You can specify the following test-options in the TEST statement.

**AGREE**

requests asymptotic tests for the simple kappa coefficient and the weighted kappa coefficient. See the sections “Simple Kappa Coefficient” on page 193 and “Weighted Kappa Coefficient” on page 194 for details.

The AGREE option in the TABLES statement provides estimates, standard errors, and confidence limits for kappa coefficients. You can request exact tests for kappa coefficients by using the EXACT statement.

Kappa coefficients are defined only for square tables, where the number of rows equals the number of columns. Kappa coefficients are not computed for tables that are not square. For $2 \times 2$ tables, the weighted kappa coefficient is identical to the simple kappa coefficient, and PROC FREQ presents only the simple kappa coefficient.

**GAMMA**

requests an asymptotic test for the gamma statistic. See the section “Gamma” on page 160 for details. The MEASURES option in the TABLES statement provides the gamma statistic and its asymptotic standard error.
KAPPA requests an asymptotic test for the simple kappa coefficient. See the section “Simple Kappa Coefficient” on page 193 for details.

The AGREE option in the TABLES statement provides the kappa statistic, its standard error, and its confidence limits. You can request an exact test for the simple kappa coefficient by specifying the KAPPA option in the EXACT statement.

Kappa coefficients are defined only for square tables, where the number of rows equals the number of columns. PROC FREQ does not compute kappa coefficients for tables that are not square.

KENTB

requests an asymptotic test for Kendall’s tau-b. See the section “Kendall’s Tau-b” on page 160 for details.

The MEASURES option in the TABLES statement provides Kendall’s tau-b and its standard error. You can request an exact test for Kendall’s tau-b by specifying the KENTB option in the EXACT statement.

MEASURES

requests asymptotic tests for the following measures of association: gamma, Kendall’s tau-b, Pearson correlation coefficient, Somers’ $D(C \mid R)$, Somers’ $D(R \mid C)$, Spearman correlation coefficient, and Stuart’s tau-c. See the section “Measures of Association” on page 158 for details.

The MEASURES option in the TABLES statement provides measures of association and their asymptotic standard errors. You can request exact tests for selected measures by using the EXACT statement.

PCORR

requests an asymptotic test for the Pearson correlation coefficient. See the section “Pearson Correlation Coefficient” on page 162 for details.

The MEASURES option in the TABLES statement provides the Pearson correlation and its standard error. You can request an exact test for the Pearson correlation by specifying the PCORR option in the EXACT statement.

PLCORR

requests Wald and likelihood ratio tests for the polychoric correlation coefficient. See the section “Polychoric Correlation” on page 164 for details.

The PLCORR option in the TABLES statement provides the polychoric correlation and its standard error.

SCORR

requests an asymptotic test for the Spearman correlation coefficient. See the section “Spearman Rank Correlation Coefficient” on page 163 for details.

The MEASURES option in the TABLES statement provides the Spearman correlation and its standard error. You can request an exact test for the Spearman correlation by specifying the SCORR option in the EXACT statement.
SMDCR requests an asymptotic test for Somers’ $D(C|R)$. See the section “Somers’ $D$” on page 161 for details.

The MEASURES option in the TABLES statement provides Somers’ $D(C|R)$ and its standard error. You can request an exact test for Somers’ $D(C|R)$ by specifying the SMDCR option in the EXACT statement.

SMDRC requests an asymptotic test for Somers’ $D(R|C)$. See the section “Somers’ $D$” on page 161 for details.

The MEASURES option in the TABLES statement provides Somers’ $D(R|C)$ and its standard error. You can request an exact test for Somers’ $D(R|C)$ by specifying the SMDRC option in the EXACT statement.

STUTC TAUC requests an asymptotic test for Stuart’s $\tau_c$. See the section “Stuart’s Tau-c” on page 161 for details.

The MEASURES option in the TABLES statement provides Stuart’s $\tau_c$ and its standard error. You can request an exact test for Stuart’s $\tau_c$ by specifying the STUTC option in the EXACT statement.

WTKAP WTKAPPA requests an asymptotic test for the weighted kappa coefficient. See the section “Weighted Kappa Coefficient” on page 194 for details.

The AGREE option in the TABLES statement provides the weighted kappa coefficient, its standard error, and confidence limits. You can request an exact test for the weighted kappa by specifying the WTKAP option in the EXACT statement.

Kappa coefficients are defined only for square tables, where the number of rows equals the number of columns. PROC FREQ does not compute kappa coefficients for tables that are not square. For $2 \times 2$ tables, the weighted kappa coefficient is identical to the simple kappa coefficient, and PROC FREQ presents only the simple kappa coefficient.

WEIGHT Statement

```
WEIGHT variable < / option > ;
```

The WEIGHT statement names a numeric variable that provides a weight for each observation in the input data set. The WEIGHT statement is most commonly used to input cell count data. See the section “Inputting Frequency Counts” on page 146 for more information. If you use a WEIGHT statement, PROC FREQ assumes that an observation represents $n$ observations, where $n$ is the value of variable. The value of the WEIGHT variable is not required to be an integer.

If the value of the WEIGHT variable is missing, PROC FREQ does not use that observation in the analysis. If the value of the WEIGHT variable is zero, PROC FREQ ignores the observation unless you specify the ZEROS option, which includes observations with zero weights. If you do not specify a WEIGHT statement, PROC FREQ assigns a weight of one to each observation. The sum of the WEIGHT variable values represents the total number of observations.
If any value of the WEIGHT variable is negative, PROC FREQ displays the frequencies computed from the weighted values but does not compute percentages and statistics. If you create an output data set by using the OUT= option in the TABLES statement, PROC FREQ assigns missing values to the PERCENT variable. PROC FREQ also assigns missing values to the variables that the OUTEXPECT and OUTPCT options provide. If any value of the WEIGHT variable is negative, you cannot create an output data set by using the OUTPUT statement because statistics are not computed when there are negative weights.

You can specify the following option in the WEIGHT statement:

**ZEROS**

includes observations with zero weight values. By default, PROC FREQ ignores observations with zero weights.

If you specify the ZEROS option, frequency and and crosstabulation tables display any levels corresponding to observations with zero weights. Without the ZEROS option, PROC FREQ does not process observations with zero weights, and so does not display levels that contain only observations with zero weights.

With the ZEROS option, PROC FREQ includes levels with zero weights in the chi-square goodness-of-fit test for one-way tables. Also, PROC FREQ includes any levels with zero weights in binomial computations for one-way tables. This makes it possible to compute binomial tests and estimates when the specified level contains no observations with positive weights.

For two-way tables, the ZEROS option enables computation of kappa statistics when there are levels that contain no observations with positive weight. For more information, see the section “Tables with Zero Rows and Columns” on page 196.

Note that even with the ZEROS option, PROC FREQ does not compute the CHISQ or MEASURES statistics for two-way tables when the table has a zero row or zero column because most of these statistics are undefined in this case.

The ZEROS option invokes the SPARSE option in the TABLES statement, which includes table cells with a zero frequency count in the LIST output and in the OUT= data set. By default, without the SPARSE option, PROC FREQ does not include zero frequency cells in the LIST output or in the OUT= data set. If you specify the ZEROS option in the WEIGHT statement but do not want the SPARSE option, you can specify the NOSPARSE option in the TABLES statement.

---

**Details: FREQ Procedure**

**Inputting Frequency Counts**

PROC FREQ can use either raw data or cell count data to produce frequency and crosstabulation tables. *Raw data*, also known as case-record data, report the data as one record for each subject or sample member. *Cell count data* report the data as a table, listing all possible combinations of data values along with the frequency counts. This way of presenting data often appears in published results.

The following DATA step statements store raw data in a SAS data set:
data Raw;
    input Subject $ R C @@;
datalines;
  01 1 1 02 1 1 03 1 1 04 1 1 05 1 1
  06 1 2 07 1 2 08 1 2 09 2 1 10 2 1
  11 2 1 12 2 1 13 2 2 14 2 2 14 2 2
;
You can store the same data as cell counts by using the following DATA step statements:

data CellCounts;
    input R C Count @@;
datalines;
  1 1 5 1 2 3
  2 1 4 2 2 3
;
The variable R contains the values for the rows, and the variable C contains the values for the columns. The variable Count contains the cell count for each row and column combination.

Both the Raw data set and the CellCounts data set produce identical frequency counts, two-way tables, and statistics. When using the CellCounts data set, you must include a WEIGHT statement to specify that the variable Count contains cell counts. For example, the following PROC FREQ statements create a two-way crosstabulation table by using the CellCounts data set:

proc freq data=CellCounts;
    tables R*C;
    weight Count;
run;

---

**Grouping with Formats**

PROC FREQ groups a variable's values according to its formatted values. If you assign a format to a variable with a FORMAT statement, PROC FREQ formats the variable values before dividing observations into the levels of a frequency or crosstabulation table.

For example, suppose that variable X has the values 1.1, 1.4, 1.7, 2.1, and 2.3. Each of these values appears as a level in the frequency table. If you decide to round each value to a single digit, include the following statement in the PROC FREQ step:

    format X 1.;

Now the table lists the frequency count for formatted level 1 as two and for formatted level 2 as three.

PROC FREQ treats formatted character variables in the same way. The formatted values are used to group the observations into the levels of a frequency table or crosstabulation table. PROC FREQ uses the entire value of a character format to classify an observation.

You can also use the FORMAT statement to assign formats that were created with the FORMAT procedure to the variables. User-written formats determine the number of levels for a variable and provide labels for a
table. If you use the same data with different formats, then you can produce frequency counts and statistics for different classifications of the variable values.

When you use PROC FORMAT to create a user-written format that combines missing and nonmissing values into one category, PROC FREQ treats the entire category of formatted values as missing. For example, a questionnaire codes 1 as yes, 2 as no, and 8 as a no answer. The following PROC FORMAT statements create a user-written format:

```sas
proc format;
  value Questfmt 1 = 'Yes'
                    2 = 'No'
                    8,. = 'Missing';
run;
```

When you use a FORMAT statement to assign `Questfmt.` to a variable, the variable’s frequency table no longer includes a frequency count for the response of 8. You must use the MISSING or MISSPRINT option in the TABLES statement to list the frequency for no answer. The frequency count for this level includes observations with either a value of 8 or a missing value (.).

The frequency or crosstabulation table lists the values of both character and numeric variables in ascending order based on internal (unformatted) variable values unless you change the order with the ORDER= option. To list the values in ascending order by formatted values, use ORDER=FORMATTED in the PROC FREQ statement.

For more information about the FORMAT statement, see *SAS Formats and Informats: Reference*.

### Missing Values

When the value of the WEIGHT variable is missing, PROC FREQ does not include that observation in the analysis.

PROC FREQ treats missing BY variable values like any other BY variable value. The missing values form a separate BY group.

If an observation has a missing value for a variable in a TABLES request, by default PROC FREQ does not include that observation in the frequency or crosstabulation table. Also by default, PROC FREQ does not include observations with missing values in the computation of percentages and statistics. The procedure displays the number of missing observations below each table.

PROC FREQ also reports the number of missing values in output data sets. The TABLES statement OUT= data set includes an observation that contains the missing value frequency. The NMISS option in the OUTPUT statement provides an output data set variable that contains the missing value frequency.

The following options change the way in which PROC FREQ handles missing values of TABLES variables:

- **MISSPRINT** displays missing value frequencies in frequency or crosstabulation tables but does not include them in computations of percentages or statistics.

- **MISSING** treats missing values as a valid nonmissing level for all TABLES variables. Displays missing levels in frequency and crosstabulation tables and includes them in computations of percentages and statistics.
This example shows the three ways that PROC FREQ can handle missing values of TABLES variables. The following DATA step statements create a data set with a missing value for the variable A:

```latex
\begin{verbatim}
data one;
    input A Freq;
datalines;
1 2
2 2
. 2
;
\end{verbatim}
```

The following PROC FREQ statements request a one-way frequency table for the variable A. The first request does not specify a missing value option. The second request specifies the MISSPRINT option in the TABLES statement. The third request specifies the MISSING option in the TABLES statement.

```latex
\begin{verbatim}
proc freq data=one;
tables A;
weight Freq;
title 'Default';
run;
proc freq data=one;
tables A / missprint;
weight Freq;
title 'MISSPRINT Option';
run;
proc freq data=one;
tables A / missing;
weight Freq;
title 'MISSING Option';
run;
\end{verbatim}
```

Figure 3.12 displays the frequency tables produced by this example. The first table shows PROC FREQ’s default behavior for handling missing values. The observation with a missing value of the TABLES variable A is not included in the table, and the frequency of missing values is displayed below the table. The second table, for which the MISSPRINT option is specified, displays the missing observation but does not include its frequency when computing the total frequency and percentages. The third table shows that PROC FREQ treats the missing level as a valid nonmissing level when the MISSING option is specified. The table displays the missing level, and PROC FREQ includes this level when computing frequencies and percentages.

**Figure 3.12** Missing Values in Frequency Tables

<table>
<thead>
<tr>
<th>A</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>50.00</td>
<td>2</td>
<td>50.00</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>50.00</td>
<td>4</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Frequency Missing = 2
When a combination of variable values for a two-way table is missing, PROC FREQ assigns zero to the frequency count for the table cell. By default, PROC FREQ does not display missing combinations in LIST format. Also, PROC FREQ does not include missing combinations in the OUT= output data set by default. To include missing combinations, you can specify the SPARSE option with the LIST or OUT= option in the TABLES statement.

In-Database Computation

The FREQ procedure can use in-database computation to construct frequency and crosstabulation tables when the DATA= input data set is stored as a table in a supported database management system (DBMS). Supported databases include Teradata, DB2 under UNIX, and Oracle. In-database computation can provide the advantages of faster processing and reduced data transfer between the database and SAS software. For information about in-database computation, see the section “In-Database Procedures” in *SAS/ACCESS for Relational Databases: Reference*.

PROC FREQ performs in-database computation by using SQL implicit pass-through. The procedure generates SQL queries that are based on the tables that you request in the TABLES statement. The database executes these SQL queries to construct initial summary tables, which are then transmitted to PROC FREQ. The procedure uses this summary information to perform the remaining analyses and tasks in the usual way (out of the database). So instead of transferring the entire data set over the network between the database and SAS software, the in-database method transfers only the summary tables. This can substantially reduce processing time when the dimensions of the summary tables (in terms of rows and columns) are much smaller than...
the dimensions of the entire database table (in terms of individual observations). Additionally, in-database summarization uses efficient parallel processing, which can also provide performance advantages.

In-database computation is controlled by the SQLGENERATION option, which you can specify in either a LIBNAME statement or an OPTIONS statement. See the section “In-Database Procedures” in SAS/ACCESS for Relational Databases: Reference for details about the SQLGENERATION option and other options that affect in-database computation. By default, PROC FREQ uses in-database computation when possible. There are no FREQ procedure options that control in-database computation.

PROC FREQ uses formatted values to group observations into the levels of frequency and crosstabulation tables. See the section “Grouping with Formats” on page 147 for more information. If formats are available in the database, then in-database summarization uses the formats. If formats are not available in the database, then in-database summarization is based on the raw data values, and PROC FREQ performs the final, formatted classification (out of the database). For more information, see the section “Deploying and Using SAS Formats in Teradata” in SAS/ACCESS for Relational Databases: Reference.

The order of observations is not inherently defined for DBMS tables. The following options relate to the order of observations and therefore should not be specified for PROC FREQ in-database computation:

- If you specify the FIRSTOBS= or OBS= data set option, PROC FREQ does not perform in-database computation.
- If you specify the NOTSORTED option in the BY statement, PROC FREQ in-database computation ignores it and uses the default ASCENDING order for BY variables.
- If you specify the ORDER=DATA option for input data in a DBMS table, PROC FREQ computation might produce different results for separate runs of the same analysis. In addition to determining the order of variable levels in crosstabulation table displays, the ORDER= option can also affect the values of many of the test statistics and measures that PROC FREQ computes.

---

**Statistical Computations**

**Definitions and Notation**

A two-way table represents the crosstabulation of row variable X and column variable Y. Let the table row values or levels be denoted by \( X_i, i = 1, 2, \ldots, R \), and the column values by \( Y_j, j = 1, 2, \ldots, C \). Let \( n_{ij} \) denote the frequency of the table cell in the \( i \)th row and \( j \)th column and define the following notation:

\[
\begin{align*}
    n_{i.} &= \sum_j n_{ij} \quad \text{(row totals)} \\
    n_{.j} &= \sum_i n_{ij} \quad \text{(column totals)} \\
    n &= \sum_i \sum_j n_{ij} \quad \text{(overall total)} \\
    p_{ij} &= n_{ij} / n \quad \text{(cell percentages)} \\
    p_{i.} &= n_{i.} / n \quad \text{(row percentages of total)} \\
    p_{.j} &= n_{.j} / n \quad \text{(column percentages of total)}
\end{align*}
\]
Chapter 3: The FREQ Procedure

\[ R_i = \text{score for row } i \]
\[ C_j = \text{score for column } j \]

\[ \bar{R} = \sum_i n_i R_i / n \quad \text{(average row score)} \]
\[ \bar{C} = \sum_j n_j C_j / n \quad \text{(average column score)} \]

\[ A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl} \]
\[ D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl} \]

\[ P = \sum_i \sum_j n_{ij} A_{ij} \quad \text{(twice the number of concordances)} \]
\[ Q = \sum_i \sum_j n_{ij} D_{ij} \quad \text{(twice the number of discordances)} \]

**Scores**

PROC FREQ uses scores of the variable values to compute the Mantel-Haenszel chi-square, Pearson correlation, Cochran-Armitage test for trend, weighted kappa coefficient, and Cochran-Mantel-Haenszel statistics. The SCORES= option in the TABLES statement specifies the score type that PROC FREQ uses. The available score types are TABLE, RANK, RIDIT, and MODRIDIT scores. The default score type is TABLE. Using MODRIDIT, RANK, or RIDIT scores yields nonparametric analyses.

For numeric variables, table scores are the values of the row and column levels. If the row or column variable is formatted, then the table score is the internal numeric value corresponding to that level. If two or more numeric values are classified into the same formatted level, then the internal numeric value for that level is the smallest of these values. For character variables, table scores are defined as the row numbers and column numbers (that is, 1 for the first row, 2 for the second row, and so on).

Rank scores, which you request with the SCORES=RANK option, are defined as

\[ R_i^1 = \sum_{k<i} n_{ki} + (n_{i+} + 1)/2 \quad i = 1, 2, \ldots, R \]
\[ C_j^1 = \sum_{l<j} n_{lj} + (n_{j+} + 1)/2 \quad j = 1, 2, \ldots, C \]

where \( R_i^1 \) is the rank score of row \( i \), and \( C_j^1 \) is the rank score of column \( j \). Note that rank scores yield midranks for tied values.

Ridit scores, which you request with the SCORES=RIDIT option, are defined as rank scores standardized by the sample size (Bross 1958; Mack and Skillings 1980). Ridit scores are derived from the rank scores as

\[ R_i^2 = R_i^1 / n \quad i = 1, 2, \ldots, R \]
\[ C_j^2 = C_j^1 / n \quad j = 1, 2, \ldots, C \]

Modified ridit scores (SCORES=MODRIDIT) represent the expected values of the order statistics of the uniform distribution on (0,1) (van Elteren 1960; Lehmann and D’Abrera 2006). Modified ridit scores are
Chi-Square Tests and Statistics

The CHISQ option provides chi-square tests of homogeneity or independence and measures of association that are based on the chi-square statistic. When you specify the CHISQ option in the TABLES statement, PROC FREQ computes the following chi-square tests for each two-way table: Pearson chi-square, likelihood ratio chi-square, and Mantel-Haenszel chi-square tests. PROC FREQ provides the following measures of association that are based on the Pearson chi-square statistic: phi coefficient, contingency coefficient, and Cramér’s V. For $2 \times 2$ tables, the CHISQ option also provides Fisher’s exact test and the continuity-adjusted chi-square statistic. You can request Fisher’s exact test for general $R \times C$ tables by specifying the FISHER option in the TABLES or EXACT statement.

If you specify the CHISQ option for one-way tables, PROC FREQ provides a one-way Pearson chi-square goodness-of-fit test. If you specify the CHISQ(LRCHI) option for one-way tables, PROC FREQ also provides a one-way likelihood ratio chi-square test. The other tests and statistics that the CHISQ option produces are available only for two-way tables.

For two-way tables, the null hypothesis for the chi-square tests is no association between the row variable and the column variable. When the sample size $n$ is large, the test statistics have asymptotic chi-square distributions under the null hypothesis. When the sample size is not large, or when the data set is sparse or heavily tied, exact tests might be more appropriate than asymptotic tests. PROC FREQ provides exact $p$-values for the Pearson chi-square, likelihood ratio chi-square, and Mantel-Haenszel chi-square tests, in addition to Fisher’s exact test. For one-way tables, PROC FREQ provides exact $p$-values for the Pearson and likelihood ratio chi-square goodness-of-fit tests. You can request these exact tests by specifying the corresponding options in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

The Mantel-Haenszel chi-square statistic is appropriate only when both variables lie on an ordinal scale. The other chi-square tests and statistics in this section are appropriate for either nominal or ordinal variables. The following sections give the formulas that PROC FREQ uses to compute the chi-square tests and statistics. For more information about these statistics, see Agresti (2007) and Stokes, Davis, and Koch (2012), and the other references cited.

Chi-Square Test for One-Way Tables

For one-way frequency tables, the CHISQ option in the TABLES statement provides a chi-square goodness-of-fit test. Let $C$ denote the number of classes, or levels, in the one-way table. Let $f_i$ denote the frequency of class $i$ (or the number of observations in class $i$) for $i = 1, 2, \ldots, C$. Then PROC FREQ computes the one-way chi-square statistic as

$$Q_P = \sum_{i=1}^{C} (f_i - e_i)^2 / e_i$$

where $e_i$ is the expected frequency for class $i$ under the null hypothesis.

In the test for equal proportions, which is the default for the CHISQ option, the null hypothesis specifies equal proportions of the total sample size for each class. Under this null hypothesis, the expected frequency
for each class equals the total sample size divided by the number of classes,

\[ e_i = \frac{n}{C} \quad \text{for} \quad i = 1, 2, \ldots, C \]

In the test for specified frequencies, which PROC FREQ computes when you input null hypothesis frequencies by using the TESTF= option, the expected frequencies are the TESTF= values that you specify. In the test for specified proportions, which PROC FREQ computes when you input null hypothesis proportions by using the TESTP= option, the expected frequencies are determined from the specified TESTP= proportions \( p_i \) as

\[ e_i = p_i \times n \quad \text{for} \quad i = 1, 2, \ldots, C \]

Under the null hypothesis (of equal proportions, specified frequencies, or specified proportions), \( Q_P \) has an asymptotic chi-square distribution with \( C-1 \) degrees of freedom.

In addition to the asymptotic test, you can request an exact one-way chi-square test by specifying the CHISQ option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Pearson Chi-Square Test for Two-Way Tables**
The Pearson chi-square for two-way tables involves the differences between the observed and expected frequencies, where the expected frequencies are computed under the null hypothesis of independence. The Pearson chi-square statistic is computed as

\[ Q_P = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \]

where \( n_{ij} \) is the observed frequency in table cell \((i, j)\) and \( e_{ij} \) is the expected frequency for table cell \((i, j)\). The expected frequency is computed under the null hypothesis that the row and column variables are independent,

\[ e_{ij} = \frac{(n_i \times n_j)}{n} \]

When the row and column variables are independent, \( Q_P \) has an asymptotic chi-square distribution with \((R-1)(C-1)\) degrees of freedom. For large values of \( Q_P \), this test rejects the null hypothesis in favor of the alternative hypothesis of general association.

In addition to the asymptotic test, you can request an exact Pearson chi-square test by specifying the PCHI or CHISQ option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

For \( 2 \times 2 \) tables, the Pearson chi-square is also appropriate for testing the equality of two binomial proportions. For \( R \times 2 \) and \( 2 \times C \) tables, the Pearson chi-square tests the homogeneity of proportions. See Fienberg (1980) for details.

**Standardized Residuals**
When you specify the CROSSLIST(STDRES) option in the TABLES statement for two-way or multiway tables, PROC FREQ displays the standardized residuals in the CROSSLIST table.

The standardized residual of a crosstabulation table cell is the ratio of \((\text{frequency} - \text{expected})\) to its standard error, where frequency is the table cell frequency and expected is the estimated expected cell frequency. The expected frequency is computed under the null hypothesis that the row and column variables are independent. See the section “Pearson Chi-Square Test for Two-Way Tables” on page 154 for more information.
PROC FREQ computes the standardized residual of table cell \((i, j)\) as

\[
\frac{(n_{ij} - e_{ij})}{\sqrt{e_{ij}(1 - p_i)(1 - p_j)}}
\]

where \(n_{ij}\) is the observed frequency of table cell \((i, j)\), \(e_{ij}\) is the expected frequency of the table cell, \(p_i\) is the proportion in row \(i\) \((n_i/n)\), and \(p_j\) is the proportion in column \(j\) \((n_j/n)\). The expected frequency of table cell \((i, j)\) is computed as

\[
e_{ij} = \frac{(n_i \cdot n_j)}{n}
\]

Under the null hypothesis of independence, each standardized residual has an asymptotic standard normal distribution. See section 2.4.5 of Agresti (2007) for more information.

**Likelihood Ratio Chi-Square Test for One-Way Tables**

For one-way frequency tables, the CHISQ(LRCHI) option in the TABLES statement provides a likelihood ratio chi-square goodness-of-fit test. By default, the likelihood ratio test is based on the null hypothesis of equal proportions in the \(C\) classes (levels) of the one-way table. If you specify null hypothesis proportions or frequencies by using the CHISQ(TESTP=) or CHISQ(TESTF=) option, respectively, the likelihood ratio test is based on the null hypothesis values that you specify.

PROC FREQ computes the one-way likelihood ratio test as

\[
G^2 = 2 \sum_{i=1}^{C} f_i \ln\left(\frac{f_i}{e_i}\right)
\]

where \(f_i\) is the observed frequency of class \(i\), and \(e_i\) is the expected frequency of class \(i\) under the null hypothesis.

For the null hypothesis of equal proportions, the expected frequency of each class equals the total sample size divided by the number of classes,

\[
e_i = \frac{n}{C} \quad \text{for} \quad i = 1, 2, \ldots, C
\]

If you provide null hypothesis frequencies by specifying the CHISQ(TESTF=) option in the TABLES statement, the expected frequencies are the TESTF= values that you specify. If you provide null hypothesis proportions by specifying the CHISQ(TESTP=) option in the TABLES statement, PROC FREQ computes the expected frequencies as

\[
e_i = p_i \times n \quad \text{for} \quad i = 1, 2, \ldots, C
\]

where the proportions \(p_i\) are the TESTP= values that you specify.

Under the null hypothesis (of equal proportions, specified frequencies, or specified proportions), the likelihood ratio statistic \(G^2\) has an asymptotic chi-square distribution with \(C-1\) degrees of freedom.

In addition to the asymptotic test, you can request an exact one-way likelihood ratio chi-square test by specifying the LRCHI option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.
**Likelihood Ratio Chi-Square Test**
The likelihood ratio chi-square involves the ratios between the observed and expected frequencies. The likelihood ratio chi-square statistic is computed as

\[ G^2 = 2 \sum_i \sum_j n_{ij} \ln \left( \frac{n_{ij}}{e_{ij}} \right) \]

where \( n_{ij} \) is the observed frequency in table cell \((i, j)\) and \( e_{ij} \) is the expected frequency for table cell \((i, j)\).

When the row and column variables are independent, \( G^2 \) has an asymptotic chi-square distribution with \((R-1)(C-1)\) degrees of freedom.

In addition to the asymptotic test, you can request an exact likelihood ratio chi-square test by specifying the LRCHI or CHISQ option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Continuity-Adjusted Chi-Square Test**
The continuity-adjusted chi-square for \(2 \times 2\) tables is similar to the Pearson chi-square, but it is adjusted for the continuity of the chi-square distribution. The continuity-adjusted chi-square is most useful for small sample sizes. The use of the continuity adjustment is somewhat controversial; this chi-square test is more conservative (and more like Fisher’s exact test) when the sample size is small. As the sample size increases, the continuity-adjusted chi-square becomes more like the Pearson chi-square.

The continuity-adjusted chi-square statistic is computed as

\[ Q_C = \sum_i \sum_j \left( \frac{\max(0, |n_{ij} - e_{ij}| - 0.5)}{e_{ij}} \right)^2 / e_{ij} \]

Under the null hypothesis of independence, \( Q_C \) has an asymptotic chi-square distribution with \((R-1)(C-1)\) degrees of freedom.

**Mantel-Haenszel Chi-Square Test**
The Mantel-Haenszel chi-square statistic tests the alternative hypothesis that there is a linear association between the row variable and the column variable. Both variables must lie on an ordinal scale. The Mantel-Haenszel chi-square statistic is computed as

\[ Q_{MH} = (n - 1)r^2 \]

where \( r \) is the Pearson correlation between the row variable and the column variable. For a description of the Pearson correlation, see the “Pearson Correlation Coefficient” on page 162. The Pearson correlation and thus the Mantel-Haenszel chi-square statistic use the scores that you specify in the SCORES= option in the TABLES statement. See Mantel and Haenszel (1959) and Landis, Heyman, and Koch (1978) for more information.

Under the null hypothesis of no association, \( Q_{MH} \) has an asymptotic chi-square distribution with one degree of freedom.

In addition to the asymptotic test, you can request an exact Mantel-Haenszel chi-square test by specifying the MHCHI or CHISQ option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.
**Fisher’s Exact Test**

Fisher’s exact test is another test of association between the row and column variables. This test assumes that the row and column totals are fixed, and then uses the hypergeometric distribution to compute probabilities of possible tables conditional on the observed row and column totals. Fisher’s exact test does not depend on any large-sample distribution assumptions, and so it is appropriate even for small sample sizes and for sparse tables.

**2 × 2 Tables**  For 2 × 2 tables, PROC FREQ gives the following information for Fisher’s exact test: table probability, two-sided \( p \)-value, left-sided \( p \)-value, and right-sided \( p \)-value. The table probability equals the hypergeometric probability of the observed table, and is in fact the value of the test statistic for Fisher’s exact test.

Where \( p \) is the hypergeometric probability of a specific table with the observed row and column totals, Fisher’s exact \( p \)-values are computed by summing probabilities \( p \) over defined sets of tables,

\[
\text{Prob} = \sum_A p
\]

The two-sided \( p \)-value is the sum of all possible table probabilities (conditional on the observed row and column totals) that are less than or equal to the observed table probability. For the two-sided \( p \)-value, the set \( A \) includes all possible tables with hypergeometric probabilities less than or equal to the probability of the observed table. A small two-sided \( p \)-value supports the alternative hypothesis of association between the row and column variables.

For 2 × 2 tables, one-sided \( p \)-values for Fisher’s exact test are defined in terms of the frequency of the cell in the first row and first column of the table, the (1,1) cell. Denoting the observed (1,1) cell frequency by \( n_{11} \), the left-sided \( p \)-value for Fisher’s exact test is the probability that the (1,1) cell frequency is less than or equal to \( n_{11} \). For the left-sided \( p \)-value, the set \( A \) includes those tables with a (1,1) cell frequency less than or equal to \( n_{11} \). A small left-sided \( p \)-value supports the alternative hypothesis that the probability of an observation being in the first cell is actually less than expected under the null hypothesis of independent row and column variables.

Similarly, for a right-sided alternative hypothesis, \( A \) is the set of tables where the frequency of the (1,1) cell is greater than or equal to that in the observed table. A small right-sided \( p \)-value supports the alternative that the probability of the first cell is actually greater than that expected under the null hypothesis.

Because the (1,1) cell frequency completely determines the 2 × 2 table when the marginal row and column sums are fixed, these one-sided alternatives can be stated equivalently in terms of other cell probabilities or ratios of cell probabilities. The left-sided alternative is equivalent to an odds ratio less than 1, where the odds ratio equals \( (n_{11}n_{22}/n_{12}n_{21}) \). Additionally, the left-sided alternative is equivalent to the column 1 risk for row 1 being less than the column 1 risk for row 2, \( p_{1|1} < p_{1|2} \). Similarly, the right-sided alternative is equivalent to the column 1 risk for row 1 being greater than the column 1 risk for row 2, \( p_{1|1} > p_{1|2} \). See Agresti (2007) for details.

**R × C Tables**  Fisher’s exact test was extended to general \( R \times C \) tables by Freeman and Halton (1951), and this test is also known as the Freeman-Halton test. For \( R \times C \) tables, the two-sided \( p \)-value definition is the same as for 2 × 2 tables. The set \( A \) contains all tables with \( p \) less than or equal to the probability of the observed table. A small \( p \)-value supports the alternative hypothesis of association between the row and column variables. For \( R \times C \) tables, Fisher’s exact test is inherently two-sided. The alternative hypothesis is defined only in terms of general, and not linear, association. Therefore, Fisher’s exact test does not have right-sided or left-sided \( p \)-values for general \( R \times C \) tables.
For $R \times C$ tables, PROC FREQ computes Fisher’s exact test by using the network algorithm of Mehta and Patel (1983), which provides a faster and more efficient solution than direct enumeration. See the section “Exact Statistics” on page 206 for more details.

**Phi Coefficient**
The phi coefficient is a measure of association derived from the Pearson chi-square. The range of the phi coefficient is $-1 \leq \phi \leq 1$ for $2 \times 2$ tables. For tables larger than $2 \times 2$, the range is $0 \leq \phi \leq \min(\sqrt{R-1}, \sqrt{C-1})$ (Liebetrau 1983). The phi coefficient is computed as

$$
\phi = \frac{(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_1n_2n_1n_2}} \quad \text{for } 2 \times 2 \text{ tables}
$$

$$
\phi = \sqrt{\frac{Q_P}{n}} \quad \text{otherwise}
$$

See Fleiss, Levin, and Paik (2003, pp. 98–99) for more information.

**Contingency Coefficient**
The contingency coefficient is a measure of association derived from the Pearson chi-square. The range of the contingency coefficient is $0 \leq P \leq \sqrt{(m-1)/m}$, where $m = \min(R, C)$ (Liebetrau 1983). The contingency coefficient is computed as

$$
P = \frac{\sqrt{Q_P}}{(Q_P + n)}
$$

See Kendall and Stuart (1979, pp. 587–588) for more information.

**Cramér’s V**
Cramér’s $V$ is a measure of association derived from the Pearson chi-square. It is designed so that the attainable upper bound is always 1. The range of Cramér’s $V$ is $-1 \leq V \leq 1$ for $2 \times 2$ tables; for tables larger than $2 \times 2$, the range is $0 \leq V \leq 1$. Cramér’s $V$ is computed as

$$
V = \phi \quad \text{for } 2 \times 2 \text{ tables}
$$

$$
V = \sqrt{\frac{Q_P/n}{\min(R-1, C-1)}} \quad \text{otherwise}
$$

See Kendall and Stuart (1979, p. 588) for more information.

**Measures of Association**

When you specify the MEASURES option in the TABLES statement, PROC FREQ computes several statistics that describe the association between the row and column variables of the contingency table. The following are measures of ordinal association that consider whether the column variable $Y$ tends to increase as the row variable $X$ increases: gamma, Kendall’s tau-$b$, Stuart’s tau-$c$, and Somers’ $D$. These measures are appropriate for ordinal variables, and they classify pairs of observations as **concordant** or **discordant**. A pair is concordant if the observation with the larger value of $X$ also has the larger value of $Y$. A pair is discordant if the observation with the larger value of $X$ has the smaller value of $Y$. See Agresti (2007) and the other references cited for the individual measures of association.

The Pearson correlation coefficient and the Spearman rank correlation coefficient are also appropriate for ordinal variables. The Pearson correlation describes the strength of the linear association between the row and column variables, and it is computed by using the row and column scores specified by the SCORES= option in the TABLES statement. The Spearman correlation is computed with rank scores. The polychoric correlation (requested by the PLCORR option) also requires ordinal variables and assumes that the variables
have an underlying bivariate normal distribution. The following measures of association do not require ordinal variables and are appropriate for nominal variables: lambda asymmetric, lambda symmetric, and the uncertainty coefficients.

PROC FREQ computes estimates of the measures according to the formulas given in the following sections. For each measure, PROC FREQ computes an asymptotic standard error (ASE), which is the square root of the asymptotic variance denoted by Var in the following sections.

**Confidence Limits**

If you specify the CL option in the TABLES statement, PROC FREQ computes asymptotic confidence limits for all MEASURES statistics. The confidence coefficient is determined according to the value of the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence limits.

The confidence limits are computed as

\[
\text{Est} \pm \left( z_{\alpha/2} \times \text{ASE} \right)
\]

where Est is the estimate of the measure, \( z_{\alpha/2} \) is the 100(1 − \( \alpha \)/2) percentile of the standard normal distribution, and ASE is the asymptotic standard error of the estimate.

**Asymptotic Tests**

For each measure that you specify in the TEST statement, PROC FREQ computes an asymptotic test of the null hypothesis that the measure equals zero. Asymptotic tests are available for the following measures of association: gamma, Kendall’s tau-\( b \), Stuart’s tau-\( c \), Somers’ \( D(C|R) \), Somers’ \( D(R|C) \), the Pearson correlation coefficient, and the Spearman rank correlation coefficient. To compute an asymptotic test, PROC FREQ uses a standardized test statistic \( z \), which has an asymptotic standard normal distribution under the null hypothesis. The test statistic is computed as

\[
z = \frac{\text{Est}}{\sqrt{\text{Var}_0(\text{Est})}}
\]

where Est is the estimate of the measure and \( \sqrt{\text{Var}_0(\text{Est})} \) is the variance of the estimate under the null hypothesis. Formulas for \( \sqrt{\text{Var}_0(\text{Est})} \) for the individual measures of association are given in the following sections.

Note that the ratio of Est to \( \sqrt{\text{Var}_0(\text{Est})} \) is the same for the following measures: gamma, Kendall’s tau-\( b \), Stuart’s tau-\( c \), Somers’ \( D(C|R) \), and Somers’ \( D(R|C) \). Therefore, the tests for these measures are identical. For example, the \( p \)-values for the test of \( H_0: \text{gamma} = 0 \) equal the \( p \)-values for the test of \( H_0: \text{tau} - b = 0 \).

PROC FREQ computes one-sided and two-sided \( p \)-values for each of these tests. When the test statistic \( z \) is greater than its null hypothesis expected value of zero, PROC FREQ displays the right-sided \( p \)-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided \( p \)-value supports the alternative hypothesis that the true value of the measure is greater than zero. When the test statistic is less than or equal to zero, PROC FREQ displays the left-sided \( p \)-value, which is the probability of a smaller value of the statistic occurring under the null hypothesis. A small left-sided \( p \)-value supports the alternative hypothesis that the true value of the measure is less than zero. The one-sided \( p \)-value \( P_1 \) can be expressed as

\[
P_1 = \begin{cases} 
\text{Prob}(Z > z) & \text{if } z > 0 \\
\text{Prob}(Z < z) & \text{if } z \leq 0 
\end{cases}
\]

where \( Z \) has a standard normal distribution. The two-sided \( p \)-value \( P_2 \) is computed as

\[
P_2 = \text{Prob}(|Z| > |z|)
\]
**Exact Tests**

Exact tests are available for the following measures of association: Kendall’s tau-b, Stuart’s tau-c, Somers’ $D(C \mid R)$ and $(R \mid C)$, the Pearson correlation coefficient, and the Spearman rank correlation coefficient. If you request an exact test for a measure of association in the EXACT statement, PROC FREQ computes the exact test of the hypothesis that the measure equals zero. See the section “Exact Statistics” on page 206 for details.

**Gamma**

The gamma ($\Gamma$) statistic is based only on the number of concordant and discordant pairs of observations. It ignores tied pairs (that is, pairs of observations that have equal values of $X$ or equal values of $Y$). Gamma is appropriate only when both variables lie on an ordinal scale. The range of gamma is $-1 \leq \Gamma \leq 1$. If the row and column variables are independent, then gamma tends to be close to zero. Gamma is computed as

$$ G = (P - Q) / (P + Q) $$

and the asymptotic variance is

$$ \text{Var}(G) = \frac{16}{(P + Q)^4} \sum_i \sum_j n_{ij} (Q A_{ij} - P D_{ij})^2 $$

For $2 \times 2$ tables, gamma is equivalent to Yule’s $Q$. See Goodman and Kruskal (1979) and Agresti (2002) for more information.

The variance under the null hypothesis that gamma equals zero is computed as

$$ \text{Var}_0(G) = \frac{4}{(P + Q)^2} \left( \sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P - Q)^2 / n \right) $$

See Brown and Benedetti (1977) for details.

**Kendall’s Tau-b**

Kendall’s tau-b ($\tau_b$) is similar to gamma except that tau-b uses a correction for ties. Tau-b is appropriate only when both variables lie on an ordinal scale. The range of tau-b is $-1 \leq \tau_b \leq 1$. Kendall’s tau-b is computed as

$$ t_b = (P - Q) / \sqrt{w_r w_c} $$

and the asymptotic variance is

$$ \text{Var}(t_b) = \frac{1}{w^4} \left( \sum_i \sum_j n_{ij} (2w d_{ij} + t_b v_{ij})^2 - n^3 t_b^2 (w_r + w_c)^2 \right) $$

where

$$ w = \sqrt{w_r w_c} $$

$$ w_r = n^2 - \sum_i n_i^2 $$

$$ w_c = n^2 - \sum_j n_j^2 $$

$$ d_{ij} = A_{ij} - D_{ij} $$

$$ v_{ij} = n_i w_c + n_j w_r $$
See Kendall (1955) for more information.

The variance under the null hypothesis that tau-$b$ equals zero is computed as

$$\text{Var}_0(t_b) = \frac{4}{w_r w_c} \left( \sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P - Q)^2 / n \right)$$

See Brown and Benedetti (1977) for details.

PROC FREQ also provides an exact test for the Kendall’s tau-$b$. You can request this test by specifying the KENTB option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Stuart’s Tau-c**

Stuart’s tau-$c$ ($\tau_c$) makes an adjustment for table size in addition to a correction for ties. Tau-$c$ is appropriate only when both variables lie on an ordinal scale. The range of tau-$c$ is $-1 \leq \tau_c \leq 1$. Stuart’s tau-$c$ is computed as

$$\tau_c = m(P - Q) / n^2(m - 1)$$

and the asymptotic variance is

$$\text{Var}(\tau_c) = \frac{4m^2}{(m - 1)^2 n^4} \left( \sum_i \sum_j n_{ij} d_{ij}^2 - (P - Q)^2 / n \right)$$

where $m = \min(R, C)$ and $d_{ij} = A_{ij} - D_{ij}$. The variance under the null hypothesis that tau-$c$ equals zero is the same as the asymptotic variance $\text{Var}$,

$$\text{Var}_0(\tau_c) = \text{Var}(\tau_c)$$

See Brown and Benedetti (1977) for details.

PROC FREQ also provides an exact test for the Stuart’s tau-$c$. You can request this test by specifying the STUTC option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Somers’ D**

Somers’ $D(C|R)$ and Somers’ $D(R|C)$ are asymmetric modifications of tau-$b$. $C|R$ indicates that the row variable $X$ is regarded as the independent variable and the column variable $Y$ is regarded as dependent. Similarly, $R|C$ indicates that the column variable $Y$ is regarded as the independent variable and the row variable $X$ is regarded as dependent. Somers’ $D$ differs from tau-$b$ in that it uses a correction only for pairs that are tied on the independent variable. Somers’ $D$ is appropriate only when both variables lie on an ordinal scale. The range of Somers’ $D$ is $-1 \leq D \leq 1$. Somers’ $D(C|R)$ is computed as

$$D(C|R) = (P - Q) / w_r$$

and its asymptotic variance is

$$\text{Var}(D(C|R)) = \frac{4}{w_r^4} \sum_i \sum_j n_{ij} (w_r d_{ij} - (P - Q)(n - n_i))^2$$

where $d_{ij} = A_{ij} - D_{ij}$ and

$$w_r = n^2 - \sum_i n_i^2.$$
For more information, see Somers (1962); Goodman and Kruskal (1979); Liebetrau (1983).

The variance under the null hypothesis that $D(C|R)$ equals zero is computed as

$$\text{Var}_0(D(C|R)) = \frac{4}{w^2} \left( \sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - \frac{(P - Q)^2}{n} \right)$$

See Brown and Benedetti (1977) for details.

Formulas for Somers’ $D(R|C)$ are obtained by interchanging the indices.

PROC FREQ also provides exact tests for Somers’ $D(C|R)$ and $(R|C)$. You can request these tests by specifying the SMDCR and SMDCR options in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Pearson Correlation Coefficient**

The Pearson correlation coefficient ($\rho$) is computed by using the scores specified in the SCORES= option. This measure is appropriate only when both variables lie on an ordinal scale. The range of the Pearson correlation is $-1 \leq \rho \leq 1$. The Pearson correlation coefficient is computed as

$$r = \frac{v}{w} = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}$$

and its asymptotic variance is

$$\text{Var}(r) = \frac{1}{w^4} \sum_i \sum_j n_{ij} \left( \frac{w(R_i - \bar{R})(C_j - \bar{C}) - \frac{b_{ij}v}{2w}}{2w} \right)^2$$

where $R_i$ and $C_j$ are the row and column scores and

$$ss_r = \sum_i \sum_j n_{ij} (R_i - \bar{R})^2$$

$$ss_c = \sum_i \sum_j n_{ij} (C_j - \bar{C})^2$$

$$ss_{rc} = \sum_i \sum_j n_{ij} (R_i - \bar{R})(C_j - \bar{C})$$

$$b_{ij} = (R_i - \bar{R})^2 ss_c + (C_j - \bar{C})^2 ss_r$$

$$v = ss_{rc}$$

$$w = \sqrt{ss_r ss_c}$$

See Snedecor and Cochran (1989) for more information.

The SCORES= option in the TABLES statement determines the type of row and column scores used to compute the Pearson correlation (and other score-based statistics). The default is SCORES=TABLE. See the section “Scores” on page 152 for details about the available score types and how they are computed.
The variance under the null hypothesis that the correlation equals zero is computed as

\[ \text{Var}_0(r) = \left( \sum_i \sum_j n_{ij}(R_i - \bar{R})(C_j - \bar{C})^2 - ss_{rc}^2/n \right) / ss_r ss_c \]

Note that this expression for the variance is derived for multinomial sampling in a contingency table framework, and it differs from the form obtained under the assumption that both variables are continuous and normally distributed. See Brown and Benedetti (1977) for details.

PROC FREQ also provides an exact test for the Pearson correlation coefficient. You can request this test by specifying the PCORR option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Spearman Rank Correlation Coefficient**

The Spearman correlation coefficient \((\rho_s)\) is computed by using rank scores, which are defined in the section “Scores” on page 152. This measure is appropriate only when both variables lie on an ordinal scale. The range of the Spearman correlation is \([-1 \leq \rho_s \leq 1]\). The Spearman correlation coefficient is computed as

\[ r_s = v / w \]

and its asymptotic variance is

\[ \text{Var}(r_s) = \frac{1}{n^2 w^4} \sum_i \sum_j n_{ij}(z_{ij} - \bar{z})^2 \]

where \(R^1_i\) and \(C^1_j\) are the row and column rank scores and

\[ v = \sum_i \sum_j n_{ij} R(i) C(j) \]

\[ w = \frac{1}{12} \sqrt{FG} \]

\[ F = n^3 - \sum_i n_i^3 \]

\[ G = n^3 - \sum_j n_j^3 \]

\[ R(i) = R^1_i - n/2 \]

\[ C(j) = C^1_j - n/2 \]

\[ \bar{z} = \frac{1}{n} \sum_i \sum_j n_{ij} z_{ij} \]

\[ z_{ij} = w v_{ij} - v w_{ij} \]
\[ v_{ij} = n \left( R(i)C(j) + \frac{1}{2} \sum_{l} n_{il}C(l) + \frac{1}{2} \sum_{k} n_{kj}R(k) + \sum_{l} \sum_{k > l} n_{kl}C(l) + \sum_{k} \sum_{i > j} n_{kl}R(k) \right) \]

\[ w_{ij} = \frac{-n}{96w^2} \left( F_{n_j^2} + G_{n_i^2} \right) \]

See Snedecor and Cochran (1989) for more information.

The variance under the null hypothesis that the correlation equals zero is computed as

\[ \text{Var}_0(r_s) = \frac{1}{n^2w^2} \sum_i \sum_j n_{ij} (v_{ij} - \bar{v})^2 \]

where

\[ \bar{v} = \sum_i \sum_j n_{ij} v_{ij} / n \]

Note that the asymptotic variance is derived for multinomial sampling in a contingency table framework, and it differs from the form obtained under the assumption that both variables are continuous and normally distributed. See Brown and Benedetti (1977) for details.

PROC FREQ also provides an exact test for the Spearman correlation coefficient. You can request this test by specifying the SCORR option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Polychoric Correlation**

When you specify the PLCORR option in the TABLES statement, PROC FREQ computes the polychoric correlation and its standard error. The polychoric correlation is based on the assumption that the two ordinal, categorical variables of the frequency table have an underlying bivariate normal distribution. The polychoric correlation coefficient is the maximum likelihood estimate of the product-moment correlation between the underlying normal variables. The range of the polychoric correlation is from –1 to 1. For \( 2 \times 2 \) tables, the polychoric correlation is also known as the tetrachoric correlation (and it is labeled as such in the displayed output). See Drasgow (1986) for an overview of polychoric correlation coefficient.

Olsson (1979) gives the likelihood equations and the asymptotic standard errors for estimating the polychoric correlation. The underlying continuous variables relate to the observed crosstabulation table through thresholds, which define a range of numeric values that correspond to each categorical (table) level. PROC FREQ uses Olsson’s maximum likelihood method for simultaneous estimation of the polychoric correlation and the thresholds. (Olsson also presents a two-step method that estimates the thresholds first.)

PROC FREQ iteratively solves the likelihood equations by using a Newton-Raphson algorithm. The initial estimates of the thresholds are computed from the inverse of the normal distribution function at the cumulative marginal proportions of the table. Iterative computation of the polychoric correlation stops when the convergence measure falls below the convergence criterion or when the maximum number of iterations is reached, whichever occurs first. For parameter values that are less than 0.01, the procedure evaluates convergence by using the absolute difference instead of the relative difference. The PLCORR(CONVERGE=)
option specifies the convergence criterion, which is 0.0001 by default. The PLCORR(MAXITER=) option specifies the maximum number of iterations, which is 20 by default.

If you specify the CL option in the TABLES statement, PROC FREQ provides confidence limits for the polychoric correlation. The confidence limits are computed as

$$\hat{\rho} \pm (z_{\alpha/2} \times SE(\hat{\rho}))$$

where \(\hat{\rho}\) is the estimate of the polychoric correlation, \(z_{\alpha/2}\) is the 100(1 - \(\alpha/2\)) percentile of the standard normal distribution, and SE(\(\hat{\rho}\)) is the standard error of the polychoric correlation estimate.

If you specify the PLCORR option in the TEST statement, PROC FREQ provides Wald and likelihood ratio tests of the null hypothesis that the polychoric correlation equals 0. The Wald test statistic is computed as

$$z = \frac{\hat{\rho}}{SE(\hat{\rho})}$$

which has a standard normal distribution under the null hypothesis. PROC FREQ computes one-sided and two-sided \(p\)-values for the Wald test. When the test statistic \(z\) is greater than its null expected value of 0, PROC FREQ displays the right-sided \(p\)-value. When the test statistic is less than or equal to 0, PROC FREQ displays the left-sided \(p\)-value.

The likelihood ratio statistic for the polychoric correlation is computed as

$$G^2 = -2 \ln(L_0/L_1)$$

where \(L_0\) is the value of the likelihood function (Olsson 1979) when the polychoric correlation is 0, and \(L_1\) is the value of the likelihood function at the maximum (where all parameters are replaced by their maximum likelihood estimates). Under the null hypothesis, the likelihood ratio statistic has an asymptotic chi-square distribution with one degree of freedom.

**Lambda (Asymmetric)**

Asymmetric lambda, \(\lambda(C \mid R)\), is interpreted as the probable improvement in predicting the column variable \(Y\) given knowledge of the row variable \(X\). The range of asymmetric lambda is \(0 \leq \lambda(C \mid R) \leq 1\). Asymmetric lambda \((C \mid R)\) is computed as

$$\lambda(C \mid R) = \frac{\sum_i r_i - r}{n - r}$$

and its asymptotic variance is

$$\text{Var}(\lambda(C \mid R)) = \frac{n - \sum_i r_i}{(n - r)^3} \left( \sum_i r_i + r - 2 \sum_i (r_i \mid r_i = l) \right)$$

where

\[
\begin{align*}
    r_i &= \max_j (n_{ij}) \\
    r &= \max_j (n_{.j}) \\
    c_j &= \max_i (n_{ij}) \\
    c &= \max_i (n_{i.})
\end{align*}
\]
The values of $l_i$ and $l$ are determined as follows. Denote by $l_i$ the unique value of $j$ such that $r_i = n_{ij}$, and let $l$ be the unique value of $j$ such that $r = n_{.j}$. Because of the uniqueness assumptions, ties in the frequencies or in the marginal totals must be broken in an arbitrary but consistent manner. In case of ties, $l_i$ is defined as the smallest value of $j$ such that $n_{ij} = r_i = c_j$, and if $cs_j = -1$, then $l_i$ is defined to be the smallest value of $j$ such that $n_{ij} = r_i = c_j$. Otherwise, if $n_{il} = r_i$, then $l_i$ is defined to be equal to $l$. If neither condition is true, then $l_i$ is taken to be the smallest value of $j$ such that $n_{ij} = r_i$.

For those columns containing a cell $(i, j)$ for which $n_{ij} = r_i = c_j$, $cs_j$ records the row in which $c_j$ is assumed to occur. Initially $cs_j$ is set equal to $-1$ for all $j$. Beginning with $i = 1$, if there is at least one value $j$ such that $n_{ij} = r_i = c_j$, and if $cs_j = 1$, then $l_i$ is defined to be the smallest such value of $j$, and $cs_j$ is set equal to $i$. Otherwise, if $n_{il} = r_i$, then $l_i$ is defined to be equal to $l_i$. If neither condition is true, then $l_i$ is taken to be the smallest value of $j$ such that $n_{ij} = r_i$.

The formulas for lambda asymmetric ($R|C$) can be obtained by interchanging the indices.

See Goodman and Kruskal (1979) for more information.

**Lambda (Symmetric)**

The nondirectional lambda is the average of the two asymmetric lambdas, $\lambda(C|R)$ and $\lambda(R|C)$. Its range is $0 \leq \lambda \leq 1$. Lambda symmetric is computed as

$$\lambda = \frac{\sum_i r_i + \sum_j c_j - r - c}{2n - r - c} = \frac{w - v}{w}$$

and its asymptotic variance is computed as

$$\text{Var}(\lambda) = \frac{1}{w^4} \left( wvy - 2w^2(n - \sum_i \sum_j (n_{ij} | j = l_i, i = k_j)) - 2v^2(n - n_{kl}) \right)$$

where

$$r_i = \max_j (n_{ij})$$

$$r = \max_j (n_{.j})$$

$$c_j = \max_i (n_{ij})$$

$$c = \max_i (n_{i.})$$

$$w = 2n - r - c$$

$$v = 2n - \sum_i r_i - \sum_j c_j$$

$$x = \sum_i (r_i | l_i = l) + \sum_j (c_j | k_j = k) + r_k + c_l$$

$$y = 8n - w - v - 2x$$

The definitions of $l_i$ and $l$ are given in the previous section. The values $k_j$ and $k$ are defined in a similar way for lambda asymmetric ($R|C$).

See Goodman and Kruskal (1979) for more information.
Uncertainty Coefficients (Asymmetric)
The uncertainty coefficient $U(C|R)$ measures the proportion of uncertainty (entropy) in the column variable $Y$ that is explained by the row variable $X$. Its range is $0 \leq U(C|R) \leq 1$. The uncertainty coefficient is computed as

$$U(C|R) = \frac{(H(X) + H(Y) - H(XY))}{H(Y)} = \frac{v}{w}$$

and its asymptotic variance is

$$\text{Var}(U(C|R)) = \frac{1}{n^2w^4} \sum_i \sum_j n_{ij} (H(Y) \ln \left(\frac{n_{ij}}{n_i}\right) + (H(X) - H(XY)) \ln \left(\frac{n_{ij}}{n}\right))^2$$

where

$$v = H(X) + H(Y) - H(XY)$$
$$w = H(Y)$$
$$H(X) = -\sum_i \left(\frac{n_{ij}}{n}\right) \ln \left(\frac{n_{ij}}{n}\right)$$
$$H(Y) = -\sum_j \left(\frac{n_{ij}}{n}\right) \ln \left(\frac{n_{ij}}{n}\right)$$
$$H(XY) = -\sum_i \sum_j \left(\frac{n_{ij}}{n}\right) \ln \left(\frac{n_{ij}}{n}\right)$$

The formulas for the uncertainty coefficient $U(R|C)$ can be obtained by interchanging the indices.

See Theil (1972, pp. 115–120) and Goodman and Kruskal (1979) for more information.

Uncertainty Coefficient (Symmetric)
The uncertainty coefficient $U$ is the symmetric version of the two asymmetric uncertainty coefficients. Its range is $0 \leq U \leq 1$. The uncertainty coefficient is computed as

$$U = \frac{2(H(X) + H(Y) - H(XY))}{(H(X) + H(Y))}$$

and its asymptotic variance is

$$\text{Var}(U) = 4 \sum_i \sum_j n_{ij} \left(\frac{H(XY) \ln \left(\frac{n_{ij}n_{ij}}{n^2}\right) - (H(X) + H(Y)) \ln \left(\frac{n_{ij}}{n}\right)}{n^2 (H(X) + H(Y))^4}\right)^2$$

where $H(X)$, $H(Y)$, and $H(XY)$ are defined in the previous section. See Goodman and Kruskal (1979) for more information.
Binomial Proportion

If you specify the BINOMIAL option in the TABLES statement, PROC FREQ computes the binomial proportion for one-way tables. By default, this is the proportion of observations in the first variable level that appears in the output. (You can use the LEVEL= option to specify a different level for the proportion.) The binomial proportion is computed as

\[ \hat{p} = \frac{n_1}{n} \]

where \( n_1 \) is the frequency of the first (or designated) level and \( n \) is the total frequency of the one-way table. The standard error of the binomial proportion is computed as

\[ se(\hat{p}) = \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}} \]

Binomial Confidence Limits

By default, PROC FREQ provides Wald and exact (Clopper-Pearson) confidence limits for the binomial proportion. If you do not specify any confidence limit requests with the CL= binomial-option, PROC FREQ computes Wald asymptotic confidence limits. You can also request Agresti-Coull, Jeffreys, and Wilson (score) confidence limits for the binomial proportion. For details about these binomial confidence limits, including comparisons of their performance, see Brown, Cai, and DasGupta (2001); Agresti and Coull (1998); Newcombe (1998b).

Wald Confidence Limits  
Wald asymptotic confidence limits are based on the normal approximation to the binomial distribution. PROC FREQ computes the Wald confidence limits for the binomial proportion as

\[ \hat{p} \pm \left( z_{\alpha/2} \times se(\hat{p}) \right) \]

where \( z_{\alpha/2} \) is the 100(1 - \( \alpha/2 \)) percentile of the standard normal distribution. The confidence level \( \alpha \) is determined by the ALPHAN= option, which, by default, equals 0.05 and produces 95% confidence limits.

If you specify CL=WALD(CORRECT) or the CORRECT binomial-option, PROC FREQ includes a continuity correction of \( 1/2n \) in the Wald asymptotic confidence limits. The purpose of this correction is to adjust for the difference between the normal approximation and the discrete binomial distribution. See Fleiss, Levin, and Paik (2003) for more information. The continuity-corrected Wald confidence limits for the binomial proportion are computed as

\[ \hat{p} \pm \left( z_{\alpha/2} \times se(\hat{p}) + (1/2n) \right) \]

Agresti-Coull Confidence Limits  
If you specify the CL=AGRESTICOULL binomial-option, PROC FREQ computes Agresti-Coull confidence limits for the binomial proportion as

\[ \hat{p} \pm \left( z_{\alpha/2} \times \sqrt{\hat{p} (1 - \hat{p}) / \tilde{n}} \right) \]

where

\[ \tilde{n}_1 = n_1 + \frac{z_{\alpha/2}^2}{2} \]
\[ \tilde{n} = n + \frac{z_{\alpha/2}^2}{2} \]
\[ \tilde{p} = \frac{\tilde{n}_1}{\tilde{n}} \]

The Agresti-Coull confidence interval has the same general form as the standard Wald interval but uses \( \tilde{p} \) in place of \( \hat{p} \). For \( \alpha = 0.05 \), the value of \( z_{\alpha/2} \) is close to 2, and this interval is the “add 2 successes and 2 failures” adjusted Wald interval of Agresti and Coull (1998).
Jeffreys Confidence Limits  If you specify the CL=JEFFREYS binomial-option, PROC FREQ computes Jeffreys confidence limits for the binomial proportion as

\[ (\beta(\alpha/2, n_1 + 1/2, n - n_1 + 1/2), \beta(1 - \alpha/2, n_1 + 1/2, n - n_1 + 1/2)) \]

where \( \beta(\alpha, b, c) \) is the \( \alpha \) percentile of the beta distribution with shape parameters \( b \) and \( c \). The lower confidence limit is set to 0 when \( n_1 = 0 \), and the upper confidence limit is set to 1 when \( n_1 = n \). This is an equal-tailed interval based on the noninformative Jeffreys prior for a binomial proportion. See Brown, Cai, and DasGupta (2001) for details. See Berger (1985) for information about using beta priors for inference on the binomial proportion.

Wilson (Score) Confidence Limits  If you specify the CL=WILSON binomial-option, PROC FREQ computes Wilson confidence limits for the binomial proportion. These are also known as score confidence limits (Wilson 1927). The confidence limits are based on inverting the normal test that uses the null proportion in the variance (the score test). Wilson confidence limits are the roots of

\[ |p - \hat{p}| = z_{\alpha/2} \sqrt{p(1-p)/n} \]

and are computed as

\[ \left( \hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2} \sqrt{\left( \frac{\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n}{n} \right) / \left( 1 + \frac{z^2_{\alpha/2}}{n} \right)} \right) / \left( 1 + \frac{z^2_{\alpha/2}}{n} \right) \]

If you specify CL=WILSON(CORRECT) or the CORRECT binomial-option, PROC FREQ provides continuity-corrected Wilson confidence limits, which are computed as the roots of

\[ |p - \hat{p}| - 1/2n = z_{\alpha/2} \sqrt{p(1-p)/n} \]

The Wilson interval has been shown to have better performance than the Wald interval and the exact (Clopper-Pearson) interval. For more information, see Agresti and Coull (1998); Brown, Cai, and DasGupta (2001); Newcombe (1998b).

Exact (Clopper-Pearson) Confidence Limits  Exact (Clopper-Pearson) confidence limits for the binomial proportion are constructed by inverting the equal-tailed test based on the binomial distribution. This method is attributed to Clopper and Pearson (1934). The exact confidence limits \( p_L \) and \( p_U \) satisfy the following equations, for \( n_1 = 1, 2, \ldots, n - 1 \):

\[ \sum_{x=n_1}^{n} \binom{n}{x} p_L^x (1 - p_L)^{n-x} = \alpha/2 \]

\[ \sum_{x=0}^{n_1} \binom{n}{x} p_U^x (1 - p_U)^{n-x} = \alpha/2 \]

The lower confidence limit equals 0 when \( n_1 = 0 \), and the upper confidence limit equals 1 when \( n_1 = n \).
PROC FREQ computes the exact (Clopper-Pearson) confidence limits by using the $F$ distribution as

$$p_L = \left( 1 + \frac{n - n_1 + 1}{n_1 F(\alpha/2, 2n_1, 2(n - n_1 + 1))} \right)^{-1}$$

$$p_U = \left( 1 + \frac{n - n_1}{(n_1 + 1) F(1 - \alpha/2, 2(n_1 + 1), 2(n - n_1))} \right)^{-1}$$

where $F(\alpha/2, b, c)$ is the $\alpha/2$ percentile of the $F$ distribution with $b$ and $c$ degrees of freedom. See Leemis and Trivedi (1996) for a derivation of this expression. Also see Collett (1991) for more information about exact binomial confidence limits.

Because this is a discrete problem, the confidence coefficient (coverage probability) of the exact (Clopper-Pearson) interval is not exactly $1 - \alpha$ but is at least $1 - \alpha$. Thus, this confidence interval is conservative. Unless the sample size is large, the actual coverage probability can be much larger than the target value. For more information about the performance of these confidence limits, see Agresti and Coull (1998); Brown, Cai, and DasGupta (2001); Leemis and Trivedi (1996).

**Binomial Tests**

The BINOMIAL option provides an asymptotic equality test for the binomial proportion by default. You can also specify `binomial-options` to request tests of noninferiority, superiority, and equivalence for the binomial proportion. If you specify the BINOMIAL option in the EXACT statement, PROC FREQ also computes exact $p$-values for the tests that you request with the `binomial-options`.

**Equality Test**

PROC FREQ computes an asymptotic test of the hypothesis that the binomial proportion equals $p_0$, where you can specify the value of $p_0$ with the `P=` `binomial-option`. If you do not specify a null value with `P=`, PROC FREQ uses $p_0 = 0.5$ by default. The binomial test statistic is computed as

$$z = (\hat{p} - p_0)/se$$

By default, the standard error is based on the null hypothesis proportion as

$$se = \sqrt{p_0(1 - p_0)/n}$$

If you specify the `VAR=SAMPLE` `binomial-option`, the standard error is computed from the sample proportion as

$$se = \sqrt{\hat{p}(1 - \hat{p})/n}$$

If you specify the `CORRECT` `binomial-option`, PROC FREQ includes a continuity correction in the asymptotic test statistic, towards adjusting for the difference between the normal approximation and the discrete binomial distribution. See Fleiss, Levin, and Paik (2003) for details. The continuity correction of $(1/2n)$ is subtracted from the numerator of the test statistic if $(\hat{p} - p_0)$ is positive; otherwise, the continuity correction is added to the numerator.

PROC FREQ computes one-sided and two-sided $p$-values for this test. When the test statistic $z$ is greater than zero (its expected value under the null hypothesis), PROC FREQ computes the right-sided $p$-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided $p$-value supports the alternative hypothesis that the true value of the proportion is greater than $p_0$. When
the test statistic is less than or equal to zero, PROC FREQ computes the left-sided $p$-value, which is the probability of a smaller value of the statistic occurring under the null hypothesis. A small left-sided $p$-value supports the alternative hypothesis that the true value of the proportion is less than $p_0$. The one-sided $p$-value $P_1$ can be expressed as

$$ P_1 = \begin{cases} 
\text{Prob}(Z > z) & \text{if } z > 0 \\
\text{Prob}(Z < z) & \text{if } z \leq 0 
\end{cases} $$

where $Z$ has a standard normal distribution. The two-sided $p$-value is computed as $P_2 = 2 \times P_1$.

If you specify the BINOMIAL option in the EXACT statement, PROC FREQ also computes an exact test of the null hypothesis $H_0: p = p_0$. To compute the exact test, PROC FREQ uses the binomial probability function,

$$ \text{Prob}(X = x | p_0) = \binom{n}{x} p_0^x (1 - p_0)^{(n-x)} \quad \text{for } x = 0, 1, 2, \ldots, n $$

where the variable $X$ has a binomial distribution with parameters $n$ and $p_0$. To compute the left-sided $p$-value, $\text{Prob}(X \leq n_1)$, PROC FREQ sums the binomial probabilities over $x$ from zero to $n_1$. To compute the right-sided $p$-value, $\text{Prob}(X \geq n_1)$, PROC FREQ sums the binomial probabilities over $x$ from $n_1$ to $n$. The exact one-sided $p$-value is the minimum of the left-sided and right-sided $p$-values,

$$ P_1 = \min ( \text{Prob}(X \leq n_1 | p_0), \text{Prob}(X \geq n_1 | p_0) ) $$

and the exact two-sided $p$-value is computed as $P_2 = 2 \times P_1$.

**Noninferiority Test** If you specify the NONINF binomial-option, PROC FREQ provides a noninferiority test for the binomial proportion. The null hypothesis for the noninferiority test is

$$ H_0: p - p_0 \leq -\delta $$

versus the alternative

$$ H_a: p - p_0 > -\delta $$

where $\delta$ is the noninferiority margin and $p_0$ is the null proportion. Rejection of the null hypothesis indicates that the binomial proportion is not inferior to the null value. See Chow, Shao, and Wang (2003) for more information.

You can specify the value of $\delta$ with the MARGIN= binomial-option, and you can specify $p_0$ with the P= binomial-option. By default, $\delta = 0.2$ and $p_0 = 0.5$.

PROC FREQ provides an asymptotic Wald test for noninferiority. The test statistic is computed as

$$ z = (\hat{p} - p_0^*) / se $$

where $p_0^*$ is the noninferiority limit,

$$ p_0^* = p_0 - \delta $$

By default, the standard error is computed from the sample proportion as

$$ se = \sqrt{\hat{p}(1 - \hat{p})/n} $$
If you specify the VAR=NULL binomial-option, the standard error is based on the noninferiority limit (determined by the null proportion and the margin) as

\[ se = \sqrt{\frac{p_0^*(1-p_0^*)}{n}} \]

If you specify the CORRECT binomial-option, PROC FREQ includes a continuity correction in the asymptotic test statistic \( z \). The continuity correction of \( \frac{1}{2n} \) is subtracted from the numerator of the test statistic if \((\hat{p} - p_0^*)\) is positive; otherwise, the continuity correction is added to the numerator.

The \( p \)-value for the noninferiority test is

\[ P_z = \text{Prob}(Z > z) \]

where \( Z \) has a standard normal distribution.

As part of the noninferiority analysis, PROC FREQ provides asymptotic Wald confidence limits for the binomial proportion. These confidence limits are computed as described in the section “Wald Confidence Limits” on page 168 but use the same standard error (VAR=NULL or VAR=SAMPLE) as the noninferiority test statistic \( z \). The confidence coefficient is \( 100(1 - 2\alpha)\% \) (Schuirmann 1999). By default, if you do not specify the ALPHA= option, the noninferiority confidence limits are 90\% confidence limits. You can compare the confidence limits to the noninferiority limit, \( p_0^* = p_0 - \delta \).

If you specify the BINOMIAL option in the EXACT statement, PROC FREQ provides an exact noninferiority test for the binomial proportion. The exact \( p \)-value is computed by using the binomial probability function with parameters \( p_0^* \) and \( n \).

\[ P_x = \sum_{k=n_1}^{k=n} \binom{n}{k} (p_0^*)^k (1-p_0^*)^{(n-k)} \]

See Chow, Shao, and Wang (2003, p. 116) for details. If you request exact binomial statistics, PROC FREQ also includes exact (Clopper-Pearson) confidence limits for the binomial proportion in the equivalence analysis display. See the section “Exact (Clopper-Pearson) Confidence Limits” on page 169 for details.

**Superiority Test**  If you specify the SUP binomial-option, PROC FREQ provides a superiority test for the binomial proportion. The null hypothesis for the superiority test is

\[ H_0: p - p_0 \leq \delta \]

versus the alternative

\[ H_a: p - p_0 > \delta \]

where \( \delta \) is the superiority margin and \( p_0 \) is the null proportion. Rejection of the null hypothesis indicates that the binomial proportion is superior to the null value. You can specify the value of \( \delta \) with the MARGIN= binomial-option, and you can specify the value of \( p_0 \) with the P= binomial-option. By default, \( \delta = 0.2 \) and \( p_0 = 0.5 \).

The superiority analysis is identical to the noninferiority analysis but uses a positive value of the margin \( \delta \) in the null hypothesis. The superiority limit equals \( p_0 + \delta \). The superiority computations follow those in the section “Noninferiority Test” on page 171 but replace \(-\delta\) with \( \delta \). See Chow, Shao, and Wang (2003) for more information.
**Equivalence Test**  If you specify the EQUIV binomial-option, PROC FREQ provides an equivalence test for the binomial proportion. The null hypothesis for the equivalence test is

\[ H_0: p - p_0 \leq \delta_L \quad \text{or} \quad p - p_0 \geq \delta_U \]

versus the alternative

\[ H_a: \delta_L < p - p_0 < \delta_U \]

where \( \delta_L \) is the lower margin, \( \delta_U \) is the upper margin, and \( p_0 \) is the null proportion. Rejection of the null hypothesis indicates that the binomial proportion is equivalent to the null value. See Chow, Shao, and Wang (2003) for more information.

You can specify the value of the margins \( \delta_L \) and \( \delta_U \) with the MARGIN= binomial-option. If you do not specify MARGIN=, PROC FREQ uses lower and upper margins of –0.2 and 0.2 by default. If you specify a single margin value \( \delta \), PROC FREQ uses lower and upper margins of –\( \delta \) and \( \delta \). You can specify the null proportion \( p_0 \) with the P= binomial-option. By default, \( p_0 = 0.5 \).

PROC FREQ computes two one-sided tests (TOST) for equivalence analysis (Schuirmann 1987). The TOST approach includes a right-sided test for the lower margin and a left-sided test for the upper margin. The overall \( p \)-value is taken to be the larger of the two \( p \)-values from the lower and upper tests.

For the lower margin, the asymptotic Wald test statistic is computed as

\[ z_L = (\hat{p} - p_L^*) / se \]

where the lower equivalence limit is

\[ p_L^* = p_0 + \delta_L \]

By default, the standard error is computed from the sample proportion as

\[ se = \sqrt{\hat{p}(1 - \hat{p})/n} \]

If you specify the VAR=NULL binomial-option, the standard error is based on the lower equivalence limit (determined by the null proportion and the lower margin) as

\[ se = \sqrt{p_L^*(1 - p_L^*)/n} \]

If you specify the CORRECT binomial-option, PROC FREQ includes a continuity correction in the asymptotic test statistic \( z_L \). The continuity correction of \( (1/2n) \) is subtracted from the numerator of the test statistic \( (\hat{p} - p_L^*) \) if the numerator is positive; otherwise, the continuity correction is added to the numerator.

The \( p \)-value for the lower margin test is

\[ P_{z,L} = \text{Prob}(Z > z_L) \]

The asymptotic test for the upper margin is computed similarly. The Wald test statistic is

\[ z_U = (\hat{p} - p_U^*) / se \]

where the upper equivalence limit is

\[ p_U^* = p_0 + \delta_U \]
By default, the standard error is computed from the sample proportion. If you specify the `VAR=NULL` binomial-option, the standard error is based on the upper equivalence limit as

\[
se = \sqrt{p_U^*(1 - p_U^*)/n}
\]

If you specify the CORRECT binomial-option, PROC FREQ includes a continuity correction of \((1/2n)\) in the asymptotic test statistic \(Z_U\).

The \(p\)-value for the upper margin test is

\[
P_{Z,U} = \text{Prob}(Z < z_U)
\]

Based on the two one-sided tests (TOST), the overall \(p\)-value for the test of equivalence equals the larger \(p\)-value from the lower and upper margin tests, which can be expressed as

\[
P_z = \max(P_{z,L}, P_{z,U})
\]

As part of the equivalence analysis, PROC FREQ provides asymptotic Wald confidence limits for the binomial proportion. These confidence limits are computed as described in the section “Wald Confidence Limits” on page 168, but use the same standard error (VAR=NULL or VAR=SAMPLE) as the equivalence test statistics and have a confidence coefficient of 100\((1 - 2\alpha)\%\) (Schuirmann 1999). By default, if you do not specify the ALPHA= option, the equivalence confidence limits are 90\% limits. If you specify VAR=NULL, separate standard errors are computed for the lower and upper margin tests, each based on the null proportion and the corresponding (lower or upper) margin. The confidence limits are computed by using the maximum of these two standard errors. You can compare the confidence limits to the equivalence limits, \((p_0 + \delta_L, p_0 + \delta_U)\).

If you specify the BINOMIAL option in the EXACT statement, PROC FREQ also provides an exact equivalence test by using two one-sided exact tests (TOST). The procedure computes lower and upper margin exact tests by using the binomial probability function as described in the section “Noninferiority Test” on page 171. The overall exact \(p\)-value for the equivalence test is taken to be the larger \(p\)-value from the lower and upper margin exact tests. If you request exact statistics, PROC FREQ also includes exact (Clopper-Pearson) confidence limits in the equivalence analysis display. The confidence coefficient is 100\((1 - 2\alpha)\%\) (Schuirmann 1999). See the section “Exact (Clopper-Pearson) Confidence Limits” on page 169 for details.

### Risks and Risk Differences

The RISKDIFF option in the TABLES statement provides estimates of risks (binomial proportions) and risk differences for \(2 \times 2\) tables. This analysis might be appropriate when comparing the proportion of some characteristic for two groups, where row 1 and row 2 correspond to the two groups, and the columns correspond to two possible characteristics or outcomes. For example, the row variable might be a treatment or dose, and the column variable might be the response. For more information, see Collett (1991); Fleiss, Levin, and Paik (2003); Stokes, Davis, and Koch (2012).

Let the frequencies of the \(2 \times 2\) table be represented as follows.

<table>
<thead>
<tr>
<th></th>
<th>Column 1</th>
<th>Column 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>(n_{11})</td>
<td>(n_{12})</td>
<td>(n_1)</td>
</tr>
<tr>
<td>Row 2</td>
<td>(n_{21})</td>
<td>(n_{22})</td>
<td>(n_2)</td>
</tr>
<tr>
<td>Total</td>
<td>(n_1)</td>
<td>(n_2)</td>
<td>(n)</td>
</tr>
</tbody>
</table>
By default when you specify the RISKDIFF option, PROC FREQ provides estimates of the row 1 risk (proportion), the row 2 risk, the overall risk, and the risk difference for column 1 and for column 2 of the 2 × 2 table. The risk difference is defined as the row 1 risk minus the row 2 risk. The risks are binomial proportions of their rows (row 1, row 2, or overall), and the computation of their standard errors and Wald confidence limits follow the binomial proportion computations, which are described in the section “Binomial Proportion” on page 168.

The column 1 risk for row 1 is the proportion of row 1 observations classified in column 1,
\[ \hat{p}_1 = \frac{n_{11}}{n_1}. \]
which estimates the conditional probability of the column 1 response, given the first level of the row variable.

The column 1 risk for row 2 is the proportion of row 2 observations classified in column 1,
\[ \hat{p}_2 = \frac{n_{21}}{n_2}. \]

The overall column 1 risk is the proportion of all observations classified in column 1,
\[ \hat{p} = \frac{n_{.1}}{n}. \]

The column 1 risk difference compares the risks for the two rows, and it is computed as the column 1 risk for row 1 minus the column 1 risk for row 2,
\[ \hat{d} = \hat{p}_1 - \hat{p}_2. \]

The standard error of the column 1 risk for row \( i \) is computed as
\[ \text{se}(\hat{p}_i) = \sqrt{\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}}. \]

The standard error of the overall column 1 risk is computed as
\[ \text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}. \]

Where the two rows represent independent binomial samples, the standard error of the column 1 risk difference is computed as
\[ \text{se}(\hat{d}) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_{11}} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_{21}}}. \]

The computations are similar for the column 2 risks and risk difference.

**Confidence Limits**

By default, the RISKDIFF option provides Wald asymptotic confidence limits for the risks (row 1, row 2, and overall) and the risk difference. By default, the RISKDIFF option also provides exact (Clopper-Pearson) confidence limits for the risks. You can suppress the display of this information by specifying the NORISKS riskdiff-option. You can specify riskdiff-options to request tests and other types of confidence limits for the risk difference. See the sections “Risk Difference Confidence Limits” on page 176 and “Risk Difference Tests” on page 179 for more information.

The risks are equivalent to the binomial proportions of their corresponding rows. This section describes the Wald confidence limits that are provided by default when you specify the RISKDIFF option. The BINOMIAL option provides additional confidence limit types and tests for risks (binomial proportions). See the sections “Binomial Confidence Limits” on page 168 and “Binomial Tests” on page 170 for details.
The Wald confidence limits are based on the normal approximation to the binomial distribution. PROC FREQ computes the Wald confidence limits for the risks and risk differences as

$$\text{Est} \pm \left( z_{\alpha/2} \times \text{se} (\text{Est}) \right)$$

where $\text{Est}$ is the estimate, $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution, and $\text{se}(\text{Est})$ is the standard error of the estimate. The confidence level $\alpha$ is determined by the value of the ALPHA= option; the default of ALPHA=0.05 produces 95% confidence limits.

If you specify the CORRECT riskdiff-option, PROC FREQ includes continuity corrections in the Wald confidence limits for the risks and risk differences. The purpose of a continuity correction is to adjust for the difference between the normal approximation and the binomial distribution, which is discrete. See Fleiss, Levin, and Paik (2003) for more information. The continuity-corrected Wald confidence limits are computed as

$$\text{Est} \pm \left( z_{\alpha/2} \times \text{se} (\text{Est}) + cc \right)$$

where $cc$ is the continuity correction. For the row 1 risk, $cc = (1/2n_1)$; for the row 2 risk, $cc = (1/2n_2)$; for the overall risk, $cc = (1/2n)$; and for the risk difference, $cc = ((1/n_1 + 1/n_2)/2)$. The column 1 and column 2 risks use the same continuity corrections.

By default when you specify the RISKDIFF option, PROC FREQ also provides exact (Clopper-Pearson) confidence limits for the column 1, column 2, and overall risks. These confidence limits are constructed by inverting the equal-tailed test that is based on the binomial distribution. See the section “Exact (Clopper-Pearson) Confidence Limits” on page 169 for details.

**Risk Difference Confidence Limits** You can request additional confidence limits for the risk difference by specifying the CL= riskdiff-option. Available confidence limit types include Agresti-Caffo, exact unconditional, Hauck-Anderson, Miettinen-Nurminen (score), Newcombe (hybrid-score), and Wald confidence limits. Continuity-corrected Newcombe and Wald confidence limits are also available.

The confidence coefficient for the confidence limits produced by the CL= riskdiff-option is $100(1 - \alpha)$%, where the value of $\alpha$ is determined by the ALPHA= option. The default of ALPHA=0.05 produces 95% confidence limits. This differs from the test-based confidence limits that are provided with the equivalence, noninferiority, and superiority tests, which have a confidence coefficient of $100(1 - 2\alpha)$% (Schuirmann 1999). See the section “Risk Difference Tests” on page 179 for details.

The section “Exact Unconditional Confidence Limits for the Risk Difference” on page 183 describes the computation of the exact confidence limits. The confidence limits are constructed by inverting two separate one-sided exact tests (tail method). By default, the tests are based on the unstandardized risk difference. If you specify the RISKDIFF(METHOD=SCORE) option, the tests are based on the score statistic.

The following sections describe the computation of the Agresti-Coull, Hauck-Anderson, Miettinen-Nurminen (score), Newcombe (hybrid-score), and Wald confidence limits for the risk difference.

**Agresti-Caffo Confidence Limits** The Agresti-Caffo confidence limits for the risk difference are computed as

$$\tilde{d} \pm \left( z_{\alpha/2} \times \text{se} (\tilde{d}) \right)$$

where $\tilde{d} = \tilde{p}_1 - \tilde{p}_2$, $\tilde{p}_1 = (n_{11} + 1)/(n_1 + 2)$, $\tilde{p}_2 = (n_{21} + 1)/(n_2 + 2)$,

$$\text{se}(\tilde{d}) = \sqrt{\tilde{p}_1(1 - \tilde{p}_1)/(n_1 + 2) + \tilde{p}_2(1 - \tilde{p}_2)/(n_2 + 2)}$$

and $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.
The Agresti-Caffo interval adjusts the Wald interval for the risk difference by adding a pseudo-observation of each type (success and failure) to each sample. See Agresti and Caffo (2000) and Agresti and Coull (1998) for more information.

**Hauck-Anderson Confidence Limits** The Hauck-Anderson confidence limits for the risk difference are computed as

\[ \hat{d} \pm \left( cc + z_{a/2} \times se(\hat{d}) \right) \]

where \( \hat{d} = \hat{p}_1 - \hat{p}_2 \) and \( z_{a/2} \) is the 100(1−\( \alpha \)/2) percentile of the standard normal distribution. The standard error is computed from the sample proportions as

\[ se(\hat{d}) = \sqrt{\hat{p}_1(1 - \hat{p}_1)(n_1 - 1) + \hat{p}_2(1 - \hat{p}_2)(n_2 - 1)} \]

The Hauck-Anderson continuity correction \( cc \) is computed as

\[ cc = 1 / (2 \min(n_1, n_2)) \]

See Hauck and Anderson (1986) for more information. The subsection “Hauck-Anderson Test” in the section “Noninferiority Tests” on page 180 describes the corresponding noninferiority test.

**Miettinen-Nurminen (Score) Confidence Limits** The Miettinen-Nurminen (score) confidence limits for the risk difference (Miettinen and Nurminen 1985) are computed by inverting score tests for the risk difference. A score-based test statistic for the null hypothesis that the risk difference equals \( \delta \) can be expressed as

\[ T(\delta) = (\hat{d} - \delta)/\sqrt{\Var(\delta)} \]

where \( \hat{d} \) is the observed value of the risk difference (\( \hat{p}_1 - \hat{p}_2 \)),

\[ \Var(\delta) = (n/(n - 1)) \left( \hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2 \right) \]

and \( \hat{p}_1(\delta) \) and \( \hat{p}_2(\delta) \) are the maximum likelihood estimates of the row 1 and row 2 risks (proportions) under the restriction that the risk difference is \( \delta \). For more information, see Miettinen and Nurminen (1985, pp. 215–216) and Miettinen (1985, chapter 12).

The 100(1−\( \alpha \))% confidence interval for the risk difference consists of all values of \( \delta \) for which the score test statistic \( T(\delta) \) falls in the acceptance region,

\[ \{ \delta : T(\delta) < z_{a/2} \} \]

where \( z_{a/2} \) is the 100(1−\( \alpha \)/2) percentile of the standard normal distribution. PROC FREQ finds the confidence limits by iterative computation, which stops when the iteration increment falls below the convergence criterion or when the maximum number of iterations is reached, whichever occurs first. By default, the convergence criterion is 0.00000001 and the maximum number of iterations is 100.

By default, the Miettinen-Nurminen confidence limits include the bias correction factor \( n/(n - 1) \) in the computation of \( \Var(\delta) \) (Miettinen and Nurminen 1985, p. 216). For more information, see Newcombe and Nurminen (2011). If you specify the CL=MN(CORRECT=NO) riskdiff-option, PROC FREQ does not include the bias correction factor in this computation (Mee 1984). See also Agresti (2002, p. 77). The uncorrected confidence limits are labeled as “Miettinen-Nurminen-Mee” confidence limits in the displayed output.
The maximum likelihood estimates of $p_1$ and $p_2$, subject to the constraint that the risk difference is $\delta$, are computed as

$$\hat{p}_1 = 2u \cos(w) - b/3a \quad \text{and} \quad \hat{p}_2 = \hat{p}_1 + \delta$$

where

$$w = (\pi + \cos^{-1}(v/u^3))/3$$

$$v = b^3/(3a)^3 - bc/6a^2 + d/2a$$

$$u = \text{sign}(v) \sqrt{\frac{b^2}{(3a)^2} - c/3a}$$

$$a = 1 + \theta$$

$$b = -(1 + \theta + \hat{p}_1 + \theta \hat{p}_2 + \delta(\theta + 2))$$

$$c = \delta^2 + \delta(2\hat{p}_1 + \theta + 1) + \hat{p}_1 + \theta \hat{p}_2$$

$$d = -\hat{p}_1 \delta(1 + \delta)$$

$$\theta = n_2/n_1.$$ 

For more information, see Farrington and Manning (1990, p. 1453).

**Newcombe Confidence Limits** Newcombe (hybrid-score) confidence limits for the risk difference are constructed from the Wilson score confidence limits for each of the two individual proportions. The confidence limits for the individual proportions are used in the standard error terms of the Wald confidence limits for the proportion difference. See Newcombe (1998a) and Barker et al. (2001) for more information.

Wilson score confidence limits for $p_1$ and $p_2$ are the roots of

$$|p_i - \hat{p}_i| = \frac{za/2 \sqrt{\hat{p}_i(1 - \hat{p}_i)}/n_i}. $$

for $i = 1, 2$. The confidence limits are computed as

$$\left(\hat{p}_i + \frac{z^2a/2}{2n_i} \pm \frac{za/2 \sqrt{(\hat{p}_i(1 - \hat{p}_i) + z^2a/4n_i)}/n_i}}{1 + \frac{z^2a/2}{n_i}}\right)$$

See the section “Wilson (Score) Confidence Limits” on page 169 for details.

Denote the lower and upper Wilson score confidence limits for $p_1$ as $L_1$ and $U_1$, and denote the lower and upper confidence limits for $p_2$ as $L_2$ and $U_2$. The Newcombe confidence limits for the proportion difference ($d = p_1 - p_2$) are computed as

$$d_L = (\hat{p}_1 - \hat{p}_2) - \sqrt{(\hat{p}_1 - L_1)^2 + (U_2 - \hat{p}_2)^2}$$

$$d_U = (\hat{p}_1 - \hat{p}_2) + \sqrt{(U_1 - \hat{p}_1)^2 + (\hat{p}_2 - L_2)^2}$$

If you specify the CORRECT `riskdiff-option`, PROC FREQ provides continuity-corrected Newcombe confidence limits. By including a continuity correction of $1/2n_i$, the Wilson score confidence limits for the individual proportions are computed as the roots of

$$|p_i - \hat{p}_i| - 1/2n_i. = \frac{za/2 \sqrt{p_i(1 - p_i)}/n_i}. $$

The continuity-corrected confidence limits for the individual proportions are then used to compute the proportion difference confidence limits $d_L$ and $d_U$. 
**Wald Confidence Limits**  The Wald confidence limits for the risk difference are computed as
\[ \hat{d} \pm (z_{\alpha/2} \times se(\hat{d})) \]
where \( \hat{d} = \hat{p}_1 - \hat{p}_2 \), \( z_{\alpha/2} \) is the 100\((1 - \alpha/2)\) percentile of the standard normal distribution, and the standard error is computed from the sample proportions as
\[ se(\hat{d}) = \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}. \]

If you specify the CORRECT riskdiff-option, the Wald confidence limits include a continuity correction \( cc \),
\[ \hat{d} \pm (cc + z_{\alpha/2} \times se(\hat{d})) \]
where \( cc = (1/n_1 + 1/n_2)/2 \).

The subsection “Wald Test” in the section “Noninferiority Tests” on page 180 describes the corresponding noninferiority test.

**Risk Difference Tests**
You can specify riskdiff-options to request tests of the risk (proportion) difference. You can request tests of equality, noninferiority, superiority, and equivalence for the risk difference. The test of equality is a standard Wald asymptotic test, available with or without a continuity correction. For noninferiority, superiority, and equivalence tests of the risk difference, the following test methods are provided: Wald (with and without continuity correction), Hauck-Anderson, Farrington-Manning (score), and Newcombe (with and without continuity correction). You can specify the test method with the METHOD= riskdiff-option. By default, PROC FREQ uses METHOD=WALD.

**Equality Test**  If you specify the EQUAL riskdiff-option, PROC FREQ computes a test of equality, or a test of the null hypothesis that the risk difference equals zero. For the column 1 (or 2) risk difference, this test can be expressed as \( H_0: d = 0 \) versus the alternative \( H_a: d \neq 0 \), where \( d = p_1 - p_2 \) denotes the column 1 (or 2) risk difference. PROC FREQ provides a Wald asymptotic test of equality. The test statistic is computed as
\[ z = \hat{d} / se(\hat{d}) \]
By default, the standard error is computed from the sample proportions as
\[ se(\hat{d}) = \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}. \]
If you specify the VAR=NULL riskdiff-option, the standard error is based on the null hypothesis that the row 1 and row 2 risks are equal,
\[ se(\hat{d}) = \sqrt{\hat{p}(1 - \hat{p}) \times (1/n_1 + 1/n_2)} \]
where \( \hat{p} = n_1/n \) estimates the overall column 1 risk.

If you specify the CORRECT riskdiff-option, PROC FREQ includes a continuity correction in the test statistic. If \( \hat{d} > 0 \), the continuity correction is subtracted from \( \hat{d} \) in the numerator of the test statistic; otherwise, the continuity correction is added to the numerator. The value of the continuity correction is \((1/n_1 + 1/n_2)/2\).

PROC FREQ computes one-sided and two-sided \( p \)-values for this test. When the test statistic \( z \) is greater than 0, PROC FREQ displays the right-sided \( p \)-value, which is the probability of a larger value occurring under the null hypothesis. The one-sided \( p \)-value can be expressed as
\[ P_1 = \begin{cases} \text{Prob}(Z > z) & \text{if } z > 0 \\ \text{Prob}(Z < z) & \text{if } z \leq 0 \end{cases} \]
where \( Z \) has a standard normal distribution. The two-sided \( p \)-value is computed as \( P_2 = 2 \times P_1 \).
**Noninferiority Tests**  If you specify the NONINF riskdiff-option, PROC FREQ provides a noninferiority test for the risk difference, or the difference between two proportions. The null hypothesis for the noninferiority test is

\[ H_0: p_1 - p_2 \leq -\delta \]

versus the alternative

\[ H_a: p_1 - p_2 > -\delta \]

where \( \delta \) is the noninferiority margin. Rejection of the null hypothesis indicates that the row 1 risk is not inferior to the row 2 risk. See Chow, Shao, and Wang (2003) for more information.

You can specify the value of \( \delta \) with the MARGIN= riskdiff-option. By default, \( \delta = 0.2 \). You can specify the test method with the METHOD= riskdiff-option. The following methods are available for the risk difference noninferiority analysis: Wald (with and without continuity correction), Hauck-Anderson, Farrington-Manning (score), and Newcombe (with and without continuity correction). The Wald, Hauck-Anderson, and Farrington-Manning methods provide tests and corresponding test-based confidence limits; the Newcombe method provides only confidence limits. If you do not specify METHOD=, PROC FREQ uses the Wald test by default.

The confidence coefficient for the test-based confidence limits is \( 100(1 - 2\alpha)\% \) (Schuirmann 1999). By default, if you do not specify the ALPHA= option, these are 90\% confidence limits. You can compare the confidence limits to the noninferiority limit, \( -\delta \).

The following sections describe the noninferiority analysis methods for the risk difference.

**Wald Test**  If you specify the METHOD=WALD riskdiff-option, PROC FREQ provides an asymptotic Wald test of noninferiority for the risk difference. This is also the default method. The Wald test statistic is computed as

\[ z = (\hat{d} + \delta) / se(\hat{d}) \]

where \( \hat{d} = \hat{p}_1 - \hat{p}_2 \) estimates the risk difference and \( \delta \) is the noninferiority margin.

By default, the standard error for the Wald test is computed from the sample proportions as

\[ se(\hat{d}) = \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}. \]

If you specify the VAR=NULL riskdiff-option, the standard error is based on the null hypothesis that the risk difference equals \( -\delta \) (Dunnett and Gent 1977). The standard error is computed as

\[ se(\hat{d}) = \sqrt{\hat{p}(1 - \hat{p})/n_2 + (\hat{p} - \delta)(1 - \hat{p} + \delta)/n_1}. \]

where

\[ \hat{p} = (n_{11} + n_{21} + \delta n_1)/n \]

If you specify the CORRECT riskdiff-option, the test statistic includes a continuity correction. The continuity correction is subtracted from the numerator of the test statistic if the numerator is greater than zero; otherwise, the continuity correction is added to the numerator. The value of the continuity correction is \( (1/n_1 + 1/n_2)/2 \).

The \( p \)-value for the Wald noninferiority test is \( P_z = \text{Prob}(Z > z) \), where \( Z \) has a standard normal distribution.
### Hauck-Anderson Test

If you specify the METHOD=HA `riskdiff-option`, PROC FREQ provides the Hauck-Anderson test for noninferiority. The Hauck-Anderson test statistic is computed as

\[ z = \frac{\hat{d} + \delta \pm cc}{se(\hat{d})} \]

where \( \hat{d} = \hat{p}_1 - \hat{p}_2 \) and the standard error is computed from the sample proportions as

\[ se(\hat{d}) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \]

The Hauck-Anderson continuity correction \( cc \) is computed as

\[ cc = 1 / (2 \min(n_1, n_2)) \]

The \( p \)-value for the Hauck-Anderson noninferiority test is \( P_z = \text{Prob}(Z > z) \), where \( Z \) has a standard normal distribution. See Hauck and Anderson (1986) and Schuirmann (1999) for more information.

### Farrington-Manning (Score) Test

If you specify the METHOD=FM `riskdiff-option`, PROC FREQ provides the Farrington-Manning (score) test of noninferiority for the risk difference. A score test statistic for the null hypothesis that the risk difference equals \(-\delta\) can be expressed as

\[ z = \frac{\hat{d} + \delta}{se(\hat{d})} \]

where \( \hat{d} \) is the observed value of the risk difference \((\hat{p}_1 - \hat{p}_2)\),

\[ se(\hat{d}) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \]

and \( \hat{p}_1 \) and \( \hat{p}_2 \) are the maximum likelihood estimates of the row 1 and row 2 risks (proportions) under the restriction that the risk difference is \(-\delta\). The \( p \)-value for the noninferiority test is \( P_z = \text{Prob}(Z > z) \), where \( Z \) has a standard normal distribution. For more information, see Miettinen and Nurminen (1985); Miettinen (1985); Farrington and Manning (1990); Dann and Koch (2005).

The maximum likelihood estimates of \( p_1 \) and \( p_1 \), subject to the constraint that the risk difference is \(-\delta\), are computed as

\[ \hat{p}_1 = 2u \cos(w) - b/3a \quad \text{and} \quad \hat{p}_2 = \hat{p}_1 + \delta \]

where

\[
\begin{align*}
w &= (\pi + \cos^{-1}(v/u^3))/3 \\
v &= b^3/(3a)^3 - bc/6a^2 + d/2a \\
u &= \text{sign}(v) \sqrt{b^2/(3a)^2 - c/3a} \\
a &= 1 + \theta \\
b &= -(1 + \theta + \hat{p}_1 + \theta \hat{p}_2 - \delta(\theta + 2)) \\
c &= \delta^2 - \delta(2\hat{p}_1 + \theta + 1) + \hat{p}_1 + \theta \hat{p}_2 \\
d &= \hat{p}_1 \delta(1 - \delta) \\
\theta &= n_2/n_1.
\end{align*}
\]

For more information, see Farrington and Manning (1990, p. 1453).

### Newcombe Noninferiority Analysis

If you specify the METHOD=NEWCOMBE `riskdiff-option`, PROC FREQ provides a noninferiority analysis that is based on Newcombe hybrid-score confidence limits for the
risk difference. The confidence coefficient for the confidence limits is \(100(1 - 2\alpha)\%\) (Schuirmann 1999). By default, if you do not specify the ALPHA= option, these are 90% confidence limits. You can compare the confidence limits with the noninferiority limit, \(-\delta\). If you specify the CORRECT riskdiff-option, the confidence limits includes a continuity correction. See the subsection “Newcombe Confidence Limits” in the section “Risk Difference Confidence Limits” on page 176 for more information.

**Superiority Test** If you specify the SUP riskdiff-option, PROC FREQ provides a superiority test for the risk difference. The null hypothesis is

\[
H_0: p_1 - p_2 \leq \delta
\]

versus the alternative

\[
H_a: p_1 - p_2 > \delta
\]

where \(\delta\) is the superiority margin. Rejection of the null hypothesis indicates that the row 1 proportion is superior to the row 2 proportion. You can specify the value of \(\delta\) with the MARGIN= riskdiff-option. By default, \(\delta = 0.2\).

The superiority analysis is identical to the noninferiority analysis but uses a positive value of the margin \(\delta\) in the null hypothesis. The superiority computations follow those in the section “Noninferiority Tests” on page 180 by replacing \(-\delta\) by \(\delta\). See Chow, Shao, and Wang (2003) for more information.

**Equivalence Tests** If you specify the EQUIV riskdiff-option, PROC FREQ provides an equivalence test for the risk difference, or the difference between two proportions. The null hypothesis for the equivalence test is

\[
H_0: p_1 - p_2 \leq \delta_L \quad \text{or} \quad p_1 - p_2 \geq \delta_U
\]

versus the alternative

\[
H_a: \delta_L < p_1 - p_2 < \delta_U
\]

where \(\delta_L\) is the lower margin and \(\delta_U\) is the upper margin. Rejection of the null hypothesis indicates that the two binomial proportions are equivalent. See Chow, Shao, and Wang (2003) for more information.

You can specify the value of the margins \(\delta_L\) and \(\delta_U\) with the MARGIN= riskdiff-option. If you do not specify MARGIN=, PROC FREQ uses lower and upper margins of \(-0.2\) and 0.2 by default. If you specify a single margin value \(\delta\), PROC FREQ uses lower and upper margins of \(-\delta\) and \(\delta\). You can specify the test method with the METHOD= riskdiff-option. The following methods are available for the risk difference equivalence analysis: Wald (with and without continuity correction), Hauck-Anderson, Farrington-Manning (score), and Newcombe (with and without continuity correction). The Wald, Hauck-Anderson, and Farrington-Manning methods provide tests and corresponding test-based confidence limits; the Newcombe method provides only confidence limits. If you do not specify METHOD=, PROC FREQ uses the Wald test by default.

PROC FREQ computes two one-sided tests (TOST) for equivalence analysis (Schuirmann 1987). The TOST approach includes a right-sided test for the lower margin \(\delta_L\) and a left-sided test for the upper margin \(\delta_U\). The overall \(p\)-value is taken to be the larger of the two \(p\)-values from the lower and upper tests.

The section “Noninferiority Tests” on page 180 gives details about the Wald, Hauck-Anderson, Farrington-Manning (score), and Newcombe methods for the risk difference. The lower margin equivalence test statistic
Statistical Computations

takes the same form as the noninferiority test statistic but uses the lower margin value \( \delta_L \) in place of \(-\delta\). The upper margin equivalence test statistic take the same form as the noninferiority test statistic but uses the upper margin value \( \delta_U \) in place of \(-\delta\).

The test-based confidence limits for the risk difference are computed according to the equivalence test method that you select. If you specify METHOD=WALD with VAR=NULL, or METHOD=FM, separate standard errors are computed for the lower and upper margin tests. In this case, the test-based confidence limits are computed by using the maximum of these two standard errors. These confidence limits have a confidence coefficient of \( 100(1 - 2\alpha)\% \) (Schuirmann 1999). By default, if you do not specify the ALPHA= option, these are 90\% confidence limits. You can compare the test-based confidence limits to the equivalence limits, \( (\delta_L, \delta_U) \).

**Exact Unconditional Confidence Limits for the Risk Difference**

If you specify the RISKDIFF option in the EXACT statement, PROC FREQ provides exact unconditional confidence limits for the risk difference. PROC FREQ computes the confidence limits by inverting two separate one-sided tests (tail method), where the size of each test is at most \( \alpha/2 \) and the confidence coefficient is at least \( (1 - \alpha) \). Exact conditional methods, described in the section “Exact Statistics” on page 206, do not apply to the risk difference due to the presence of a nuisance parameter (Agresti 1992). The unconditional approach eliminates the nuisance parameter by maximizing the \( p \)-value over all possible values of the parameter (Santner and Snell 1980).

By default, PROC FREQ uses the unstandardized risk difference as the test statistic in the confidence limit computations. If you specify the RISKDIFF(METHOD=SCORE) option, the procedure uses the score statistic (Chan and Zhang 1999). The score statistic is a less discrete statistic than the raw risk difference and produces less conservative confidence limits (Agresti and Min 2001). See also Santner et al. (2007) for comparisons of methods for computing exact confidence limits for the risk difference.

PROC FREQ computes the confidence limits as follows. The risk difference is defined as the difference between the row 1 and row 2 risks (proportions), \( d = p_1 - p_2 \), and \( n_1 \) and \( n_2 \) denote the row totals of the \( 2 \times 2 \) table. The joint probability function for the table can be expressed in terms of the table cell frequencies, the risk difference, and the nuisance parameter \( p_2 \) as

\[
f(n_{11}, n_{21}, n_1, n_2, d, p_2) = \binom{n_1}{n_{11}} (d + p_2)^{n_{11}} (1 - d - p_2)^{n_1 - n_{11}} \times \binom{n_2}{n_{21}} p_2^{n_{21}} (1 - p_2)^{n_2 - n_{21}}
\]

The \( 100(1 - \alpha/2)\% \) confidence limits for the risk difference are computed as

\[
d_L = \sup_{p_2} (d_* : P_U(d_*) > \alpha/2) \\
d_U = \inf_{p_2} (d_* : P_L(d_*) > \alpha/2)
\]

where

\[
P_U(d_*) = \sup_{p_2} \left( \sum_{A,T(a) \geq t_0} f(n_{11}, n_{21}, n_1, n_2, d_*, p_2) \right)
\]

\[
P_L(d_*) = \sup_{p_2} \left( \sum_{A,T(a) \leq t_0} f(n_{11}, n_{21}, n_1, n_2, d_*, p_2) \right)
\]

The set \( A \) includes all \( 2 \times 2 \) tables with row sums equal to \( n_1 \) and \( n_2 \), and \( T(a) \) denotes the value of the test statistic for table \( a \) in \( A \). To compute \( P_U(d_*) \), the sum includes probabilities of those tables for which...
Chapter 3: The FREQ Procedure

\( T(a) \geq t_0 \), where \( t_0 \) is the value of the test statistic for the observed table. For a fixed value of \( d_* \), \( P_U(d_*) \) is taken to be the maximum sum over all possible values of \( p_2 \).

By default, PROC FREQ uses the unstandardized risk difference as the test statistic \( T \). If you specify the RISKDIFF(METHOD=SCORE) option, the procedure uses the risk difference score statistic as the test statistic (Chan and Zhang 1999). For information about the computation of the score statistic, see the section “” on page 177. For more information, see Miettinen and Nurminen (1985) and Farrington and Manning (1990).

**Barnard’s Unconditional Exact Test**

The BARNARD option in the EXACT statement provides an unconditional exact test for the risk (proportion) difference for \( 2 \times 2 \) tables. The reference set for the unconditional exact test consists of all \( 2 \times 2 \) tables that have the same row sums as the observed table (Barnard 1945, 1947, 1949). This differs from the reference set for exact conditional inference, which is restricted to the set of tables that have the same row sums and the same column sums as the observed table. See the sections “Fisher’s Exact Test” on page 157 and “Exact Statistics” on page 206 for more information.

The test statistic is the standardized risk difference, which is computed as

\[
T = \frac{d}{\sqrt{p_1(1-p_1)(1/n_1 + 1/n_2)}}
\]

where the risk difference \( d \) is defined as the difference between the row 1 and row 2 risks (proportions), \( d = (n_{11}/n_1 - n_{21}/n_2) \); \( n_1 \) and \( n_2 \) are the row 1 and row 2 totals, respectively; and \( p_1 \) is the overall proportion in column 1, \( (n_{11} + n_{21})/n \).

Under the null hypothesis that the risk difference is 0, the joint probability function for a table can be expressed in terms of the table cell frequencies, the row totals, and the unknown parameter \( \pi \) as

\[
f(n_{11}, n_{21}; n_1, n_2, \pi) = \binom{n_1}{n_{11}} \binom{n_2}{n_{21}} \pi^{n_{11} + n_{21}} (1 - \pi)^{n - n_{11} - n_{21}}
\]

where \( \pi \) is the common value of the risk (proportion).

PROC FREQ sums the table probabilities over the reference set for those tables where the test statistic is greater than or equal to the observed value of the test statistic. This sum can be expressed as

\[
\Pr(\pi) = \sum_{A, T(a) \geq t_0} f(n_{11}, n_{21}; n_1, n_2, \pi)
\]

where the set \( A \) contains all \( 2 \times 2 \) tables with row sums equal to \( n_1 \) and \( n_2 \), and \( T(a) \) denotes the value of the test statistic for table \( a \) in \( A \). The sum includes probabilities of those tables for which \( (T(a) \geq t_0) \), where \( t_0 \) is the value of the test statistic for the observed table.

The sum \( \Pr(\pi) \) depends on the unknown value of \( \pi \). To compute the exact \( p \)-value, PROC FREQ eliminates the nuisance parameter \( \pi \) by taking the maximum value of \( \Pr(\pi) \) over all possible values of \( \pi \),

\[
\Pr = \sup_{0 \leq \pi \leq 1} (\Pr(\pi))
\]

Common Risk Difference

PROC FREQ provides Mantel-Haenszel and summary score estimates for the common risk (proportion) difference for multiway $2 \times 2$ tables.

**Mantel-Haenszel Estimate of the Common Risk Difference**

PROC FREQ computes the Mantel-Haenszel (Mantel and Haenszel 1959) estimate of the common risk difference as

$$\hat{d}_{MH} = \left( \frac{\sum h \hat{d}_h w_h}{\sum h \hat{w}_h} \right)$$

where $\hat{d}_h$ is the risk difference in stratum $h$ and

$$w_h = \frac{n_{h1} n_{h2}}{n_h}$$

The column 1 risk difference in stratum $h$ is computed as

$$\hat{d}_h = \hat{p}_{h1} - \hat{p}_{h2} = \left( \frac{n_{h11}}{n_{h1}}\right) - \left( \frac{n_{h21}}{n_{h2}}\right)$$

where $\hat{p}_{h1}$ is the proportion of row 1 observations that are classified in column 1 and $\hat{p}_{h2}$ is the proportion or row 2 observations that are classified in column 1. The column 2 risk is computed in the same way. For more information, see Agresti (2013, p. 231).

PROC FREQ computes the variance of $\hat{d}_{MH}$ (Sato 1989) as

$$\hat{\sigma}^2(\hat{d}_{MH}) = \left( \frac{\hat{d}_{MH} \sum h P_h + \sum h Q_h}{\sum h \hat{w}_h} \right)^2$$

where

$$P_h = \left( \frac{n_{h1}^2 n_{h21} - n_{h2}^2 n_{h11} + n_{h1} n_{h2} (n_{h2} - n_{h1})}{n_h^2} \right) / n_h$$

$$Q_h = \left( \frac{n_{h11} (n_{h2} - n_{h21}) + n_{h21} (n_{h1} - n_{h11})}{2 n_h} \right)$$

The 100(1 - $\alpha$/2)% confidence limits for the common risk difference are

$$\hat{d}_{MH} \pm \left( z_{\alpha/2} \times \hat{\sigma}(\hat{d}_{MH}) \right)$$

**Summary Score Estimate of the Common Risk Difference**

PROC FREQ computes the summary score estimate of the common risk difference as described in Agresti (2013, p. 231). This estimate is computed from Miettinen-Nurminen (score) confidence limits for the stratum risk differences. For more information, see the section “Miettinen-Nurminen (Score) Confidence Limits.”

The score confidence interval for the risk difference in stratum $h$ can be expressed as $\hat{d}_h' \pm z_{\alpha/2} \hat{s}_h'$, where $\hat{d}_h'$ is the midpoint of the score confidence interval and $\hat{s}_h'$ is the width of the confidence interval divided by $2z_{\alpha/2}$. The summary score estimate of the common risk difference is computed as

$$\hat{d}_S = \sum h \hat{d}_h w_h'$$
where

$$w'_h = (1/s'_h^2) \sum_i (1/s'_i^2)$$

The variance of $\hat{d}_S$ is computed as

$$\hat{\sigma}^2(\hat{d}_S) = 1/ \sum_h (1/s'_h^2)$$

The 100(1 $- \alpha/2$)% summary score confidence limits for the common risk difference are

$$\hat{d}_S \pm \left( z_{\alpha/2} \times \hat{\sigma}(\hat{d}_S) \right)$$

**Odds Ratio and Relative Risks for 2 x 2 Tables**

**Odds Ratio**

The odds ratio is a useful measure of association for a variety of study designs. For a retrospective design called a case-control study, the odds ratio can be used to estimate the relative risk when the probability of positive response is small (Agresti 2002). In a case-control study, two independent samples are identified based on a binary (yes-no) response variable, and the conditional distribution of a binary explanatory variable is examined, within fixed levels of the response variable. See Stokes, Davis, and Koch (2012) and Agresti (2007).

The odds of a positive response (column 1) in row 1 is $n_{11}/n_{12}$. Similarly, the odds of a positive response in row 2 is $n_{21}/n_{22}$. The odds ratio is formed as the ratio of the row 1 odds to the row 2 odds. The odds ratio for a $2 \times 2$ table is defined as

$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

The odds ratio can be any nonnegative number. When the row and column variables are independent, the true value of the odds ratio equals 1. An odds ratio greater than 1 indicates that the odds of a positive response are higher in row 1 than in row 2. Values less than 1 indicate the odds of positive response are higher in row 2. The strength of association increases with the deviation from 1.

The transformation $G = (OR - 1)/(OR + 1)$ transforms the odds ratio to the range $(-1, 1)$ with $G = 0$ when $OR = 1$; $G = -1$ when $OR = 0$; and $G$ approaches 1 as $OR$ approaches infinity. $G$ is the gamma statistic, which PROC FREQ computes when you specify the MEASURES option.

The asymptotic 100(1 $- \alpha$)% confidence limits for the odds ratio are

$$\left( OR \times \exp(-z \sqrt{v}), ~ OR \times \exp(z \sqrt{v}) \right)$$

where

$$v = \text{Var}(\ln OR) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

and $z$ is the 100(1 $- \alpha/2$) percentile of the standard normal distribution. If any of the four cell frequencies are zero, the estimates are not computed.
Score Confidence Limits for the Odds Ratio  Score confidence limits for the odds ratio (Miettinen and Nurminen 1985) are computed by inverting score tests for the odds ratio. A score-based chi-square test statistic for the null hypothesis that the odds ratio equals \( \theta \) can be expressed as

\[
Q(\theta) = \left\{ \frac{(\hat{p}_1 - \hat{p}_1)}{\frac{n}{(n - 1)}} \right\} - 1
\]

where \( \hat{p}_1 \) is the observed row 1 risk (proportion), and \( \hat{p}_1 \) and \( \hat{p}_2 \) are the maximum likelihood estimates of the row 1 and row 2 risks under the restriction that the odds ratio \( \frac{n_{11}n_{22}}{n_{12}n_{21}} \) is \( \theta \). For more information, see Miettinen and Nurminen (1985) and Miettinen (1985, chapter 14).

The 100(1 - \( \alpha \))% score confidence interval for the odds ratio consists of all values of \( \theta \) for which the test statistic \( Q(\theta) \) falls in the acceptance region,

\[ \{ \theta : Q(\theta) < \chi^2_{1,\alpha} \} \]

where \( \chi^2_{1,\alpha} \) is the 100(1 - \( \alpha \)) percentile of the chi-square distribution with 1 degree of freedom. PROC FREQ finds the confidence limits by iterative computation. For more information about score confidence limits, see Agresti (2013).

By default, the score confidence limits include the bias correction factor \( \frac{n}{(n - 1)} \) in the denominator of \( Q(\theta) \) (Miettinen and Nurminen 1985, p. 217). If you specify the CL=SCORE(CORRECT=NO) option, PROC FREQ does not include this factor in the computation.

The maximum likelihood estimates of \( p_1 \) and \( p_2 \), subject to the constraint that the odds ratio is \( \theta \), are computed as

\[
\hat{p}_2 = \left( -b + (b^2 - 4ac) \right) / 2a \quad \text{and} \quad \hat{p}_1 = \frac{\hat{p}_2\theta}{1 + \hat{p}_2(\theta - 1)}
\]

where

\[
a = n_2(\theta - 1) \\
b = n_1\theta + n_2 - \hat{p}_1(\theta - 1) \\
c = -\hat{p}_1
\]

For more information, see Miettinen and Nurminen (1985, pp. 217–218) and Miettinen (1985, chapter 14).

Exact Confidence Limits for the Odds Ratio  When you specify the OR option in the EXACT statement, PROC FREQ computes exact confidence limits for the odds ratio. Because this is a discrete problem, the confidence coefficient for the exact confidence interval is not exactly \( 1 - \alpha \) but is at least \( 1 - \alpha \). Thus, these confidence limits are conservative. See Agresti (1992) for more information.

PROC FREQ computes exact confidence limits for the odds ratio by using an algorithm based on Thomas (1971). See also Gart (1971). The following two equations are solved iteratively to determine the lower and upper confidence limits, \( \phi_1 \) and \( \phi_2 \):

\[
\sum_{i=n_{11}}^{n_1} \binom{n_1}{i} \binom{n_2}{n_1-i} \phi_1^i / \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{n_1-i} \phi_1^i = \alpha / 2
\]

\[
\sum_{i=0}^{n_{11}} \binom{n_1}{i} \binom{n_2}{n_1-i} \phi_2^i / \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{n_1-i} \phi_2^i = \alpha / 2
\]
When the odds ratio equals zero, which occurs when either \( n_{11} = 0 \) or \( n_{22} = 0 \), PROC FREQ sets the lower exact confidence limit to zero and determines the upper limit with level \( \alpha \). Similarly, when the odds ratio equals infinity, which occurs when either \( n_{12} = 0 \) or \( n_{21} = 0 \), PROC FREQ sets the upper exact confidence limit to infinity and determines the lower limit with level \( \alpha \).

**Relative Risks**

These measures of relative risk are useful in cohort (prospective) study designs, where two samples are identified based on the presence or absence of an explanatory factor. The two samples are observed in future time for the binary (yes-no) response variable under study. Relative risk measures are also useful in cross-sectional studies, where two variables are observed simultaneously. See Stokes, Davis, and Koch (2012) and Agresti (2007) for more information.

The column 1 relative risk is the ratio of the column 1 risk for row 1 to row 2. The column 1 risk for row 1 is the proportion of the row 1 observations classified in column 1,

\[
p_1 = \frac{n_{11}}{n_1}.
\]

Similarly, the column 1 risk for row 2 is

\[
p_2 = \frac{n_{21}}{n_2}.
\]

The column 1 relative risk is computed as

\[
RR_1 = \frac{p_1}{p_2}
\]

A relative risk greater than 1 indicates that the probability of positive response is greater in row 1 than in row 2. Similarly, a relative risk less than 1 indicates that the probability of positive response is less in row 1 than in row 2. The strength of association increases with the deviation from 1.

Asymptotic 100(1 – \( \alpha \))% confidence limits for the column 1 relative risk are computed as

\[
\left( RR_1 \times \exp(-z \sqrt{v}), \quad RR_1 \times \exp(z \sqrt{v}) \right)
\]

where

\[
v = \text{Var}(\ln RR_1) = \left( \frac{(1 - p_1)/n_{11}} + \frac{(1 - p_2)/n_{21}} \right)
\]

and \( z \) is the 100(1 – \( \alpha \)/2) percentile of the standard normal distribution. If either \( n_{11} \) or \( n_{21} \) is zero, the estimates are not computed.

PROC FREQ computes the column 2 relative risks in the same way.

**Exact Unconditional Confidence Limits for the Relative Risk** If you specify the RELRISK option in the EXACT statement, PROC FREQ provides exact unconditional confidence limits for the relative risk. PROC FREQ computes the confidence limits by inverting two separate one-sided tests (tail method), where the size of each test is at most \( \alpha / 2 \) and the confidence coefficient is at least \( 1 - \alpha \). Exact conditional methods, described in the section “Exact Statistics” on page 206, do not apply to the relative risk due to the presence of a nuisance parameter (Agresti 1992). The unconditional approach eliminates the nuisance parameter by maximizing the \( p \)-value over all possible values of the parameter (Santner and Snell 1980).

By default, PROC FREQ uses the unstandardized relative risk as the test statistic in the confidence limit computations. If you specify the RELRISK(METHOD=SCORE) option, the procedure uses the relative risk
score statistic (Chan and Zhang 1999). The score statistic is a less discrete statistic than the raw relative risk and produces less conservative confidence limits (Agresti and Min 2001). See also Santner et al. (2007) for comparisons of methods for computing exact confidence limits.

See the section “Exact Unconditional Confidence Limits for the Risk Difference” on page 183 for a description of the method that PROC FREQ uses to compute confidence limits for the relative risk. The test statistic for the relative risk computation is either the unstandardized relative risk (by default) or the relative risk score statistic (if you specify the RELRISK(METHOD=SCORE) option). PROC FREQ uses the following form of the unstandardized relative risk, which adds 0.05 to each frequency, to ensure that the statistic is defined when there are zero table cells (Gart and Nam 1988):

$$rr = \frac{(n_{11} + 0.5) / (n_1 + 0.5)}{(n_{21} + 0.5) / (n_2 + 0.5)}$$

If you specify the RELRISK(METHOD=SCORE) option, PROC FREQ uses the relative risk score statistic (Miettinen and Nurminen 1985; Farrington and Manning 1990). This test statistic is computed as

$$z = (\hat{p}_1 - R_0 \hat{p}_2) / \text{se}(rr)$$

where

$$\text{se}(rr) = \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + R_0^2 \hat{p}_2(1 - \hat{p}_2)/n_2}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the maximum likelihood estimators of $p_1$ and $p_2$ under the null hypothesis that the relative risk equals $R_0$. From Farrington and Manning (1990, p. 1454), the maximum likelihood solution is

$$\hat{p}_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad \hat{p}_2 = \hat{p}_1 / R_0$$

where

$$a = 1 + \theta$$
$$b = -(R_0(1 + \theta \hat{p}_2) + \theta + \hat{p}_1)$$
$$c = R_0(\hat{p}_1 + \theta \hat{p}_2)$$
$$\theta = n_2/n_1.$$  

**Cochran-Armitage Test for Trend**

The TREND option in the TABLES statement provides the Cochran-Armitage test for trend, which tests for trend in binomial proportions across levels of a single factor or covariate. This test is appropriate for a two-way table where one variable has two levels and the other variable is ordinal. The two-level variable represents the response, and the other variable represents an explanatory variable with ordered levels. When the two-way has two columns and $R$ rows, PROC FREQ tests for trend across the $R$ levels of the row variable, and the binomial proportion is computed as the proportion of observations in the first column. When the table has two rows and $C$ columns, PROC FREQ tests for trend across the $C$ levels of the column variable, and the binomial proportion is computed as the proportion of observations in the first row.

The trend test is based on the regression coefficient for the weighted linear regression of the binomial proportions on the scores of the explanatory variable levels. See Margolin (1988) and Agresti (2002) for details. If the table has two columns and $R$ rows, the trend test statistic is computed as

$$T = \sum_{i=1}^{R} n_{i1}(R_i - \bar{R}) / \sqrt{p_1 (1 - p_1) s^2}$$
where $R_i$ is the score of row $i$, $\bar{R}$ is the average row score, and

$$s^2 = \sum_{i=1}^{R} n_i (R_i - \bar{R})^2$$

The SCORES= option in the TABLES statement determines the type of row scores used in computing the trend test (and other score-based statistics). The default is SCORES=TABLE. See the section “Scores” on page 152 for details. For character variables, the table scores for the row variable are the row numbers (for example, 1 for the first row, 2 for the second row, and so on). For numeric variables, the table score for each row is the numeric value of the row level. When you perform the trend test, the explanatory variable might be numeric (for example, dose of a test substance), and the variable values might be appropriate scores. If the explanatory variable has ordinal levels that are not numeric, you can assign meaningful scores to the variable levels. Sometimes equidistant scores, such as the table scores for a character variable, might be appropriate. For more information on choosing scores for the trend test, see Margolin (1988).

The null hypothesis for the Cochran-Armitage test is no trend, which means that the binomial proportion $p_{i1} = n_{i1}/n_i$ is the same for all levels of the explanatory variable. Under the null hypothesis, the trend statistic has an asymptotic standard normal distribution.

PROC FREQ computes one-sided and two-sided $p$-values for the trend test. When the test statistic is greater than its null hypothesis expected value of zero, PROC FREQ displays the right-sided $p$-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided $p$-value supports the alternative hypothesis of increasing trend in proportions from row 1 to row $R$. When the test statistic is less than or equal to zero, PROC FREQ displays the left-sided $p$-value. A small left-sided $p$-value supports the alternative of decreasing trend.

The one-sided $p$-value for the trend test is computed as

$$P_1 = \begin{cases} \Pr(Z > T) & \text{if } T > 0 \\ \Pr(Z < T) & \text{if } T \leq 0 \end{cases}$$

where $Z$ has a standard normal distribution. The two-sided $p$-value is computed as

$$P_2 = \Pr(|Z| > |T|)$$

PROC FREQ also provides exact $p$-values for the Cochran-Armitage trend test. You can request the exact test by specifying the TREND option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Jonckheere-Terpstra Test**

The JT option in the TABLES statement provides the Jonckheere-Terpstra test, which is a nonparametric test for ordered differences among classes. It tests the null hypothesis that the distribution of the response variable does not differ among classes. It is designed to detect alternatives of ordered class differences, which can be expressed as $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_R$ (or $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_R$), with at least one of the inequalities being strict, where $\tau_i$ denotes the effect of class $i$. For such ordered alternatives, the Jonckheere-Terpstra test can be preferable to tests of more general class difference alternatives, such as the Kruskal–Wallis test (produced by the WILCOXON option in the NPAR1WAY procedure). See Pirie (1983) and Hollander and Wolfe (1999) for more information about the Jonckheere-Terpstra test.
The Jonckheere-Terpstra test is appropriate for a two-way table in which an ordinal column variable represents the response. The row variable, which can be nominal or ordinal, represents the classification variable. The levels of the row variable should be ordered according to the ordering you want the test to detect. The order of variable levels is determined by the ORDER= option in the PROC FREQ statement. The default is ORDER=INTERNAL, which orders by unformatted values. If you specify ORDER=DATA, PROC FREQ orders values according to their order in the input data set. For more information about how to order variable levels, see the ORDER= option.

The Jonckheere-Terpstra test statistic is computed by first forming $R(R - 1)/2$ Mann-Whitney counts $M_{i,i'}$, where $i < i'$, for pairs of rows in the contingency table,

\[
M_{i,i'} = \begin{cases} 
\text{number of times } X_{i,j} < X_{i',j'}, & j = 1, \ldots, n_i; \quad j' = 1, \ldots, n_{i'}; \\
+ \frac{1}{2} \{ \text{number of times } X_{i,j} = X_{i',j'}, & j = 1, \ldots, n_i; \quad j' = 1, \ldots, n_{i'}; 
\end{cases}
\]

where $X_{i,j}$ is response $j$ in row $i$. The Jonckheere-Terpstra test statistic is computed as

\[
J = \sum_{1 \leq j < i' \leq R} \sum M_{i,i'}
\]

This test rejects the null hypothesis of no difference among classes for large values of $J$. Asymptotic $p$-values for the Jonckheere-Terpstra test are obtained by using the normal approximation for the distribution of the standardized test statistic. The standardized test statistic is computed as

\[
J^* = \frac{(J - E_0(J))}{\sqrt{\text{Var}_0(J)}}
\]

where $E_0(J)$ and $\text{Var}_0(J)$ are the expected value and variance of the test statistic under the null hypothesis,

\[
E_0(J) = \left( n^2 - \sum_i n_i^2 \right) / 4
\]

\[
\text{Var}_0(J) = A/72 + B / (36n(n - 1)(n - 2)) + C / (8n(n - 1))
\]

where

\[
A = n(n - 1)(2n + 5) - \sum_i n_i(n_i - 1)(2n_i + 5) - \sum_j n_j(n_j - 1)(2n_j + 5)
\]

\[
B = \left( \sum_i n_i(n_i - 1)(n_i - 2) \right) \left( \sum_j n_j(n_j - 1)(n_j - 2) \right)
\]

\[
C = \left( \sum_i n_i(n_i - 1) \right) \left( \sum_j n_j(n_j - 1) \right)
\]

PROC FREQ computes one-sided and two-sided $p$-values for the Jonckheere-Terpstra test. When the standardized test statistic is greater than its null hypothesis expected value of zero, PROC FREQ displays the right-sided $p$-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided $p$-value supports the alternative hypothesis of increasing order from row 1 to row $R$. When the standardized test statistic is less than or equal to zero, PROC FREQ displays the left-sided $p$-value. A small left-sided $p$-value supports the alternative of decreasing order from row 1 to row $R$. 


The one-sided $p$-value for the Jonckheere-Terpstra test, $P_1$, is computed as

$$P_1 = \begin{cases} \text{Prob}(Z > J^*) & \text{if } J^* > 0 \\ \text{Prob}(Z < J^*) & \text{if } J^* \leq 0 \end{cases}$$

where $Z$ has a standard normal distribution. The two-sided $p$-value, $P_2$, is computed as

$$P_2 = \text{Prob}(|Z| > |J^*|)$$

PROC FREQ also provides exact $p$-values for the Jonckheere-Terpstra test. You can request the exact test by specifying the JT option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

Tests and Measures of Agreement

When you specify the AGREE option in the TABLES statement, PROC FREQ computes tests and measures of agreement for square tables (that is, for tables where the number of rows equals the number of columns). For two-way tables, these tests and measures include McNemar’s test for $2 \times 2$ tables, Bowker’s test of symmetry, the simple kappa coefficient, and the weighted kappa coefficient. For multiple strata ($n$-way tables, where $n > 2$), PROC FREQ also computes the overall simple kappa coefficient and the overall weighted kappa coefficient, as well as tests for equal kappas (simple and weighted) among strata. Cochran’s $Q$ is computed for multiway tables when each variable has two levels, that is, for $h \times 2 \times 2$ tables.

PROC FREQ computes the kappa coefficients (simple and weighted), their asymptotic standard errors, and their confidence limits when you specify the AGREE option in the TABLES statement. If you also specify the KAPPA option in the TEST statement, then PROC FREQ computes the asymptotic test of the hypothesis that simple kappa equals zero. Similarly, if you specify the WTKAP option in the TEST statement, PROC FREQ computes the asymptotic test for weighted kappa.

In addition to the asymptotic tests described in this section, PROC FREQ provides exact $p$-values for McNemar’s test, the simple kappa coefficient test, and the weighted kappa coefficient test. You can request these exact tests by specifying the corresponding options in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

The following sections provide the formulas that PROC FREQ uses to compute the AGREE statistics. For information about the use and interpretation of these statistics, see Agresti (2002, 2007); Fleiss, Levin, and Paik (2003), and the other references cited for each statistic.

McNemar’s Test

PROC FREQ computes McNemar’s test for $2 \times 2$ tables when you specify the AGREE option. McNemar’s test is appropriate when you are analyzing data from matched pairs of subjects with a dichotomous (yes-no) response. It tests the null hypothesis of marginal homogeneity, or $p_{11} = p_{21}$. McNemar’s test is computed as

$$Q_M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Under the null hypothesis, $Q_M$ has an asymptotic chi-square distribution with one degree of freedom. See McNemar (1947), as well as the general references cited in the preceding section. In addition to the asymptotic test, PROC FREQ also computes the exact $p$-value for McNemar’s test when you specify the MCNEM option in the EXACT statement.
**Bowker’s Test of Symmetry**

For Bowker’s test of symmetry, the null hypothesis is that the cell proportions are symmetric, or that $p_{ij} = p_{ji}$ for all pairs of table cells. For $2 \times 2$ tables, Bowker’s test is identical to McNemar’s test, and so PROC FREQ provides Bowker’s test for square tables larger than $2 \times 2$.

Bowker’s test of symmetry is computed as

$$Q_B = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})}$$

For large samples, $Q_B$ has an asymptotic chi-square distribution with $R(R - 1)/2$ degrees of freedom under the null hypothesis of symmetry. See Bowker (1948) for details.

**Simple Kappa Coefficient**

The simple kappa coefficient, introduced by Cohen (1960), is a measure of interrater agreement. PROC FREQ computes the simple kappa coefficient as

$$\hat{\kappa} = \frac{(P_o - P_e)}{(1 - P_e)}$$

where $P_o = \sum_i p_{ii}$ and $P_e = \sum_i p_i.p_i$. If the two response variables are viewed as two independent ratings of the $n$ subjects, the kappa coefficient equals +1 when there is complete agreement of the raters. When the observed agreement exceeds chance agreement, kappa is positive, with its magnitude reflecting the strength of agreement. Although this is unusual in practice, kappa is negative when the observed agreement is less than chance agreement. The minimum value of kappa is between –1 and 0, depending on the marginal proportions.

The asymptotic variance of the simple kappa coefficient is computed as

$$\text{Var}(\hat{\kappa}) = \frac{(A + B - C)}{(1 - P_e)^2 n}$$

where

$$A = \sum_i p_{ii} (1 - (p_i + p_i)(1 - \hat{\kappa}))^2$$

$$B = (1 - \hat{\kappa})^2 \sum_{i \neq j} p_{ij}(p_i + p_j)^2$$

$$C = (\hat{\kappa} - P_e(1 - \hat{\kappa}))^2$$

See Fleiss, Cohen, and Everitt (1969) for details.

PROC FREQ computes confidence limits for the simple kappa coefficient as

$$\hat{\kappa} \pm \left( z_{\alpha/2} \times \sqrt{\text{Var}(\hat{\kappa})} \right)$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution. The value of $\alpha$ is determined by the value of the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence limits.

To compute an asymptotic test for the kappa coefficient, PROC FREQ uses the standardized test statistic $\hat{\kappa}^*$, which has an asymptotic standard normal distribution under the null hypothesis that kappa equals zero. The standardized test statistic is computed as

$$\hat{\kappa}^* = \frac{\hat{\kappa}}{\sqrt{\text{Var}(\hat{\kappa})}}$$
where \( \text{Var}_0(\hat{\kappa}) \) is the variance of the kappa coefficient under the null hypothesis,

\[
\text{Var}_0(\hat{\kappa}) = \left( P_e + P_e^2 - \sum_i p_i \cdot p_{i.} (p_{i.} + p_{.i}) \right) / (1 - P_e)^2 n
\]


PROC FREQ also provides an exact test for the simple kappa coefficient. You can request the exact test by specifying the KAPPA or AGREE option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Weighted Kappa Coefficient**

The weighted kappa coefficient is a generalization of the simple kappa coefficient that uses weights to quantify the relative difference between categories. For \( 2 \times 2 \) tables, the weighted kappa coefficient equals the simple kappa coefficient. PROC FREQ displays the weighted kappa coefficient only for tables larger than \( 2 \times 2 \). PROC FREQ computes the kappa weights from the column scores, by using either Cicchetti-Allison weights or Fleiss-Cohen weights, both of which are described in the following section. The weights \( w_{ij} \) are constructed so that \( 0 \leq w_{ij} < 1 \) for all \( i \neq j \), \( w_{ii} = 1 \) for all \( i \), and \( w_{ij} = w_{ji} \). The weighted kappa coefficient is computed as

\[
\hat{\kappa}_w = \left( P_{o(w)} - P_{e(w)} \right) / (1 - P_{e(w)})
\]

where

\[
P_{o(w)} = \sum_i \sum_j w_{ij} p_{ij}
\]

\[
P_{e(w)} = \sum_i \sum_j w_{ij} p_{i.} p_{.j}
\]

The asymptotic variance of the weighted kappa coefficient is

\[
\text{Var}(\hat{\kappa}_w) = \left( \sum_i \sum_j p_{ij} \left( w_{ij} - (\overline{w}_{i.} + \overline{w}_{.j})(1 - \hat{\kappa}_w) \right)^2 - (\hat{\kappa}_w - P_{e(w)}(1 - \hat{\kappa}_w))^2 \right) / (1 - P_{e(w)})^2 n
\]

where

\[
\overline{w}_{i.} = \sum_j p_{i.} w_{ij}
\]

\[
\overline{w}_{.j} = \sum_i p_{i.} w_{ij}
\]

See Fleiss, Cohen, and Everitt (1969) for details.

PROC FREQ computes confidence limits for the weighted kappa coefficient as

\[
\hat{\kappa}_w \pm \left( z_{\alpha/2} \times \sqrt{\text{Var}(\hat{\kappa}_w)} \right)
\]

where \( z_{\alpha/2} \) is the \( 100(1 - \alpha/2) \) percentile of the standard normal distribution. The value of \( \alpha \) is determined by the value of the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence limits.
To compute an asymptotic test for the weighted kappa coefficient, PROC FREQ uses the standardized test statistic $\hat{k}_w^*$, which has an asymptotic standard normal distribution under the null hypothesis that weighted kappa equals zero. The standardized test statistic is computed as

$$\hat{k}_w^* = \frac{\hat{k}_w}{\sqrt{\text{Var}_0(\hat{k}_w)}}$$

where $\text{Var}_0(\hat{k}_w)$ is the variance of the weighted kappa coefficient under the null hypothesis,

$$\text{Var}_0(\hat{k}_w) = \left( \sum_i \sum_j p_i \cdot p_j \left( w_{ij} - (\bar{w}_i + \bar{w}_j) \right)^2 - P^2_{e(w)} \right) / (1 - P_{e(w)})^2 n$$


PROC FREQ also provides an exact test for the weighted kappa coefficient. You can request the exact test by specifying the WTKAPPA or AGREE option in the EXACT statement. See the section “Exact Statistics” on page 206 for more information.

**Weights**  PROC FREQ computes kappa coefficient weights by using the column scores and one of the two available weight types. The column scores are determined by the SCORES= option in the TABLES statement. The two available types of kappa weights are Cicchetti-Allison and Fleiss-Cohen weights. By default, PROC FREQ uses Cicchetti-Allison weights. If you specify (WT=FC) with the AGREE option, then PROC FREQ uses Fleiss-Cohen weights to compute the weighted kappa coefficient.

PROC FREQ computes Cicchetti-Allison kappa coefficient weights as

$$w_{ij} = 1 - \frac{C_i - C_j}{C_C - C_1}$$

where $C_i$ is the score for column $i$ and $C$ is the number of categories or columns. See Cicchetti and Allison (1971) for details.

The SCORES= option in the TABLES statement determines the type of column scores used to compute the kappa weights (and other score-based statistics). The default is SCORES=TABLE. See the section “Scores” on page 152 for details. For numeric variables, table scores are the values of the variable levels. You can assign numeric values to the levels in a way that reflects their level of similarity. For example, suppose you have four levels and order them according to similarity. If you assign them values of 0, 2, 4, and 10, the Cicchetti-Allison kappa weights take the following values: $w_{12} = 0.8$, $w_{13} = 0.6$, $w_{14} = 0$, $w_{23} = 0.8$, $w_{24} = 0.2$, and $w_{34} = 0.4$. Note that when there are only two categories (that is, $C = 2$), the weighted kappa coefficient is identical to the simple kappa coefficient.

If you specify (WT=FC) with the AGREE option in the TABLES statement, PROC FREQ computes Fleiss-Cohen kappa coefficient weights as

$$w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_C - C_1)^2}$$

See Fleiss and Cohen (1973) for details.

For the preceding example, the Fleiss-Cohen kappa weights are: $w_{12} = 0.96$, $w_{13} = 0.84$, $w_{14} = 0$, $w_{23} = 0.96$, $w_{24} = 0.36$, and $w_{34} = 0.64$. 


**Overall Kappa Coefficient**

When there are multiple strata, PROC FREQ combines the stratum-level estimates of kappa into an overall estimate of the supposed common value of kappa. Assume there are \( q \) strata, indexed by \( h = 1, 2, \ldots, q \), and let \( \text{Var}(\hat{k}_h) \) denote the variance of \( \hat{k}_h \). The estimate of the overall kappa coefficient is computed as

\[
\hat{k}_T = \frac{\sum_{h=1}^{q} \hat{k}_h}{\sum_{h=1}^{q} \text{Var}(\hat{k}_h)} / \frac{1}{\sum_{h=1}^{q} \text{Var}(\hat{k}_h)}
\]


PROC FREQ computes an estimate of the overall weighted kappa in the same way.

**Tests for Equal Kappa Coefficients**

When there are multiple strata, the following chi-square statistic tests whether the stratum-level values of kappa are equal:

\[
Q_K = \sum_{h=1}^{q} (\hat{k}_h - \hat{k}_T)^2 / \text{Var}(\hat{k}_h)
\]

Under the null hypothesis of equal kappas for the \( q \) strata, \( Q_K \) has an asymptotic chi-square distribution with \( q-1 \) degrees of freedom. See Fleiss, Levin, and Paik (2003) for more information. PROC FREQ computes a test for equal weighted kappa coefficients in the same way.

**Cochran’s Q Test**

Cochran’s \( Q \) is computed for multiway tables when each variable has two levels, that is, for \( 2 \times 2 \times \cdots \times 2 \) tables. Cochran’s \( Q \) statistic is used to test the homogeneity of the one-dimensional margins. Let \( m \) denote the number of variables and \( N \) denote the total number of subjects. Cochran’s \( Q \) statistic is computed as

\[
Q_C = m(m-1) \left( \frac{\sum_{j=1}^{m} T_j^2 - T^2}{m(T - \sum_{k=1}^{N} S_k^2)} \right)
\]

where \( T_j \) is the number of positive responses for variable \( j \), \( T \) is the total number of positive responses over all variables, and \( S_k \) is the number of positive responses for subject \( k \). Under the null hypothesis, Cochran’s \( Q \) has an asymptotic chi-square distribution with \( m-1 \) degrees of freedom. See Cochran (1950) for details. When there are only two binary response variables (\( m=2 \)), Cochran’s \( Q \) simplifies to McNemar’s test. When there are more than two response categories, you can test for marginal homogeneity by using the repeated measures capabilities of the CATMOD procedure.

**Tables with Zero Rows and Columns**

The AGREE statistics are defined only for square tables, where the number of rows equals the number of columns. If the table is not square, PROC FREQ does not compute AGREE statistics. In the kappa statistic framework, where two independent raters assign ratings to each of \( n \) subjects, suppose one of the raters does not use all possible \( r \) rating levels. If the corresponding table has \( r \) rows but only \( r-1 \) columns, then the table is not square and PROC FREQ does not compute AGREE statistics. To create a square table in this situation, use the ZEROS option in the WEIGHT statement, which requests that PROC FREQ include observations with zero weights in the analysis. Include zero-weight observations in the input data set to
represent any rating levels that are not used by a rater, so that the input data set has at least one observation for each possible rater and rating combination. The analysis then includes all rating levels, even when all levels are not actually assigned by both raters. The resulting table (of rater 1 by rater 2) is a square table, and AGREE statistics can be computed.

For more information, see the description of the ZEROS option. By default, PROC FREQ does not process observations that have zero weights, because these observations do not contribute to the total frequency count, and because any resulting zero-weight row or column causes many of the tests and measures of association to be undefined. However, kappa statistics are defined for tables with a zero-weight row or column, and the ZEROS option makes it possible to input zero-weight observations and construct the tables needed to compute kappas.

**Cochran-Mantel-Haenszel Statistics**

The CMH option in the TABLES statement gives a stratified statistical analysis of the relationship between the row and column variables after controlling for the strata variables in a multiway table. For example, for the table request \( A \times B \times C \times D \), the CMH option provides an analysis of the relationship between \( C \) and \( D \), after controlling for \( A \) and \( B \). The stratified analysis provides a way to adjust for the possible confounding effects of \( A \) and \( B \) without being forced to estimate parameters for them.

The CMH analysis produces Cochran-Mantel-Haenszel statistics, which include the correlation statistic, the ANOVA (row mean scores) statistic, and the general association statistic. For \( 2 \times 2 \) tables, the CMH option also provides Mantel-Haenszel and logit estimates of the common odds ratio and the common relative risks, as well as the Breslow-Day test for homogeneity of the odds ratios.

Exact statistics are also available for stratified \( 2 \times 2 \) tables. If you specify the EQOR option in the EXACT statement, PROC FREQ provides Zelen’s exact test for equal odds ratios. If you specify the COMOR option in the EXACT statement, PROC FREQ provides exact confidence limits for the common odds ratio and an exact test that the common odds ratio equals one.

Let the number of strata be denoted by \( q \), indexing the strata by \( h = 1, 2, \ldots, q \). Each stratum contains a contingency table with \( X \) representing the row variable and \( Y \) representing the column variable. For table \( h \), denote the cell frequency in row \( i \) and column \( j \) by \( n_{hi,j} \), with corresponding row and column marginal totals denoted by \( n_{hi} \) and \( n_{h,j} \), and the overall stratum total by \( n_h \).

Because the formulas for the Cochran-Mantel-Haenszel statistics are more easily defined in terms of matrices, the following notation is used. Vectors are presumed to be column vectors unless they are transposed (\(^t\)).

\[
\begin{align*}
n'_{hi} & = (n'_{hi1}, n'_{hi2}, \ldots, n'_{hiC}) \quad (1 \times C) \\
n'_h & = (n'_h1, n'_h2, \ldots, n'_hC) \quad (1 \times RC) \\
p_{hi} & = n_{hi} / n_h \quad (1 \times 1) \\
p_{h,j} & = n_{h,j} / n_h \quad (1 \times 1) \\
P'_{h*} & = (p_{h1}, p_{h2}, \ldots, p_{hR}) \quad (1 \times R) \\
P'_{*h} & = (p_{h1}, p_{h2}, \ldots, p_{hC}) \quad (1 \times C)
\end{align*}
\]

Assume that the strata are independent and that the marginal totals of each stratum are fixed. The null hypothesis, \( H_0 \), is that there is no association between \( X \) and \( Y \) in any of the strata. The corresponding model
Chapter 3: The FREQ Procedure

is the multiple hypergeometric; this implies that, under $H_0$, the expected value and covariance matrix of the frequencies are, respectively,

$$m_h = \mathbb{E}[n_h \mid H_0] = n_h (P_{h*} \otimes P_{h*})$$

$$\text{Var}[n_h \mid H_0] = c \left( (D_{P_{h*}} - P_{h*} P_{h*}') \otimes (D_{P_{h*}} - P_{h*} P_{h*}') \right)$$

where

$$c = n_h^2 / (n_h - 1)$$

and where $\otimes$ denotes Kronecker product multiplication and $D_a$ is a diagonal matrix with the elements of $a$ on the main diagonal.

The generalized CMH statistic (Landis, Heyman, and Koch 1978) is defined as

$$Q_{CMH} = G'V_G^{-1}G$$

where

$$G = \sum_h B_h (n_h - m_h)$$

$$V_G = \sum_h B_h (\text{Var}[n_h \mid H_0]) B_h'$$

and where

$$B_h = C_h \otimes R_h$$

is a matrix of fixed constants based on column scores $C_h$ and row scores $R_h$. When the null hypothesis is true, the CMH statistic has an asymptotic chi-square distribution with degrees of freedom equal to the rank of $B_h$. If $V_G$ is found to be singular, PROC FREQ prints a message and sets the value of the CMH statistic to missing.

PROC FREQ computes three CMH statistics by using this formula for the generalized CMH statistic, with different row and column score definitions for each statistic. The CMH statistics that PROC FREQ computes are the correlation statistic, the ANOVA (row mean scores) statistic, and the general association statistic. These statistics test the null hypothesis of no association against different alternative hypotheses. The following sections describe the computation of these CMH statistics.

**CAUTION:** The CMH statistics have low power for detecting an association in which the patterns of association for some of the strata are in the opposite direction of the patterns displayed by other strata. Thus, a nonsignificant CMH statistic suggests either that there is no association or that no pattern of association has enough strength or consistency to dominate any other pattern.

**Correlation Statistic**

The correlation statistic, popularized by Mantel and Haenszel, has one degree of freedom and is known as the Mantel-Haenszel statistic (Mantel and Haenszel 1959; Mantel 1963).

The alternative hypothesis for the correlation statistic is that there is a linear association between $X$ and $Y$ in at least one stratum. If either $X$ or $Y$ does not lie on an ordinal (or interval) scale, then this statistic is not meaningful.
To compute the correlation statistic, PROC FREQ uses the formula for the generalized CMH statistic with the row and column scores determined by the SCORES= option in the TABLES statement. See the section “Scores” on page 152 for more information about the available score types. The matrix of row scores $R_h$ has dimension $1 \times R$, and the matrix of column scores $C_h$ has dimension $1 \times C$.

When there is only one stratum, this CMH statistic reduces to $(n - 1) r^2$, where $r$ is the Pearson correlation coefficient between $X$ and $Y$. When nonparametric (RANK or RIDIT) scores are specified, the statistic reduces to $(n - 1) r_{s}^2$, where $r_{s}$ is the Spearman rank correlation coefficient between $X$ and $Y$. When there is more than one stratum, this CMH statistic becomes a stratum-adjusted correlation statistic.

**ANOVA (Row Mean Scores) Statistic**

The ANOVA statistic can be used only when the column variable $Y$ lies on an ordinal (or interval) scale so that the mean score of $Y$ is meaningful. For the ANOVA statistic, the mean score is computed for each row of the table, and the alternative hypothesis is that, for at least one stratum, the mean scores of the $R$ rows are unequal. In other words, the statistic is sensitive to location differences among the $R$ distributions of $Y$.

The matrix of column scores $C_h$ has dimension $1 \times C$, and the column scores are determined by the SCORES= option.

The matrix of row scores $R_h$ has dimension $(R - 1) \times R$ and is created internally by PROC FREQ as

$$R_h = [I_{R-1}, -J_{R-1}]$$

where $I_{R-1}$ is an identity matrix of rank $R - 1$ and $J_{R-1}$ is an $(R - 1) \times 1$ vector of ones. This matrix has the effect of forming $R - 1$ independent contrasts of the $R$ mean scores.

When there is only one stratum, this CMH statistic is essentially an analysis of variance (ANOVA) statistic in the sense that it is a function of the variance ratio $F$ statistic that would be obtained from a one-way ANOVA on the dependent variable $Y$. If nonparametric scores are specified in this case, then the ANOVA statistic is a Kruskal-Wallis test.

If there is more than one stratum, then this CMH statistic corresponds to a stratum-adjusted ANOVA or Kruskal-Wallis test. In the special case where there is one subject per row and one subject per column in the contingency table of each stratum, this CMH statistic is identical to Friedman’s chi-square. See Example 3.9 for an illustration.

**General Association Statistic**

The alternative hypothesis for the general association statistic is that, for at least one stratum, there is some kind of association between $X$ and $Y$. This statistic is always interpretable because it does not require an ordinal scale for either $X$ or $Y$.

For the general association statistic, the matrix $R_h$ is the same as the one used for the ANOVA statistic. The matrix $C_h$ is defined similarly as

$$C_h = [I_{C-1}, -J_{C-1}]$$

PROC FREQ generates both score matrices internally. When there is only one stratum, then the general association CMH statistic reduces to $Q_P (n - 1)/n$, where $Q_P$ is the Pearson chi-square statistic. When there is more than one stratum, then the CMH statistic becomes a stratum-adjusted Pearson chi-square statistic. Note that a similar adjustment can be made by summing the Pearson chi-squares across the strata. However,
the latter statistic requires a large sample size in each stratum to support the resulting chi-square distribution with \(q(R-1)(C-1)\) degrees of freedom. The CMH statistic requires only a large overall sample size because it has only \((R-1)(C-1)\) degrees of freedom.

See Cochran (1954); Mantel and Haenszel (1959); Mantel (1963); Birch (1965); Landis, Heyman, and Koch (1978).

**Mantel-Fleiss Criterion**

If you specify the CMH(MANTELFLEISS) option in the TABLES statement, PROC FREQ computes the Mantel-Fleiss criterion for stratified \(2 \times 2\) tables. The Mantel-Fleiss criterion can be used to assess the validity of the chi-square approximation for the distribution of the Mantel-Haenszel statistic for \(2 \times 2\) tables. For details, see Mantel and Fleiss (1980); Mantel and Haenszel (1959); Stokes, Davis, and Koch (2012); Dmitrienko et al. (2005).

The Mantel-Fleiss criterion is computed as

\[
MF = \min \left( \left[ \sum_h m_{h11} - \sum_h (n_{h11}L) \right], \left[ \sum_h (n_{h11}U) - \sum_h m_{h11} \right] \right)
\]

where \(m_{h11}\) is the expected value of \(n_{h11}\) under the hypothesis of no association between the row and column variables in table \(h\), \((n_{h11})_L\) is the minimum possible value of the table cell frequency, and \((n_{h11})_U\) is the maximum possible value,

\[
m_{h11} = n_{h1-} n_{h-1} / n_h
\]

\[
(n_{h11})_L = \max(0, n_{h1-} - n_{h2-})
\]

\[
(n_{h11})_U = \min(n_{h1-}, n_{h-1})
\]

The Mantel-Fleiss guideline accepts the validity of the Mantel-Haenszel approximation when the value of the criterion is at least 5. When the criterion is less than 5, PROC FREQ displays a warning.

**Adjusted Odds Ratio and Relative Risk Estimates**

The CMH option provides adjusted odds ratio and relative risk estimates for stratified \(2 \times 2\) tables. For each of these measures, PROC FREQ computes a Mantel-Haenszel estimate and a logit estimate. These estimates apply to \(n\)-way table requests in the TABLES statement, when the row and column variables both have two levels.

For example, for the table request \(A* B * C * D\), if the row and column variables \(C\) and \(D\) both have two levels, PROC FREQ provides odds ratio and relative risk estimates, adjusting for the confounding variables \(A\) and \(B\).

The choice of an appropriate measure depends on the study design. For case-control (retrospective) studies, the odds ratio is appropriate. For cohort (prospective) or cross-sectional studies, the relative risk is appropriate. See the section “Odds Ratio and Relative Risks for \(2 \times 2\) Tables” on page 186 for more information on these measures.

Throughout this section, \(z\) denotes the \(100(1 - \alpha/2)\) percentile of the standard normal distribution.
Odds Ratio, Case-Control Studies  
PROC FREQ provides Mantel-Haenszel and logit estimates for the common odds ratio for stratified $2 \times 2$ tables.

**Mantel-Haenszel Estimator**  
The Mantel-Haenszel estimate of the common odds ratio is computed as

$$OR_{MH} = \left( \sum_h n_{h11} n_{h22} / n_h \right) / \left( \sum_h n_{h12} n_{h21} / n_h \right)$$

It is always computed unless the denominator is zero. See Mantel and Haenszel (1959) and Agresti (2002) for details.

To compute confidence limits for the common odds ratio, PROC FREQ uses the Robins, Breslow, and Greenland (1986) variance estimate for $\ln(OR_{MH})$. The $100(1 - \alpha/2)\%$ confidence limits for the common odds ratio are

$$\left( OR_{MH} \times \exp(-z\hat{\delta}), \ OR_{MH} \times \exp(z\hat{\delta}) \right)$$

where

$$\hat{\delta}^2 = \widehat{\text{Var}}(\ln(OR_{MH}))$$

$$= \sum_h (n_{h11} + n_{h22})(n_{h11} n_{h22})/n_h^2 \left\{ 2 \left( \sum_h n_{h11} n_{h22} / n_h \right)^2 \right. $$

$$+ \left. \sum_h ((n_{h11} + n_{h22})(n_{h12} n_{h21}) + (n_{h12} + n_{h21})(n_{h11} n_{h22}))/n_h^2 \right\}$$

$$= \sum_h (n_{h12} + n_{h21})(n_{h12} n_{h21})/2 \left( \sum_h n_{h12} n_{h21} / n_h \right) \left( \sum_h n_{h12} n_{h21} / n_h \right)$$

Note that the Mantel-Haenszel odds ratio estimator is less sensitive to small $n_h$ than the logit estimator.

**Logit Estimator**  
The adjusted logit estimate of the common odds ratio (Woolf 1955) is computed as

$$OR_L = \exp \left( \sum_h w_h \ln(OR_h) / \sum_h w_h \right)$$

and the corresponding $100(1 - \alpha)\%$ confidence limits are

$$\left( OR_L \times \exp \left( -z / \sqrt{\sum_h w_h} \right), \ OR_L \times \exp \left( z / \sqrt{\sum_h w_h} \right) \right)$$

where $OR_h$ is the odds ratio for stratum $h$, and

$$w_h = 1 / \text{Var}(\ln(OR_h))$$

If any table cell frequency in a stratum $h$ is zero, PROC FREQ adds 0.5 to each cell of the stratum before computing $OR_h$ and $w_h$ (Haldane 1955) for the logit estimate. The procedure prints a warning when this occurs.
Relative Risks, Cohort Studies  PROC FREQ provides Mantel-Haenszel and logit estimates of the common relative risks for stratified 2 × 2 tables.

Mantel-Haenszel Estimator The Mantel-Haenszel estimate of the common relative risk for column 1 is computed as

\[
RR_{MH} = \frac{\left( \sum_h n_{h11} n_{h21} / n_h \right)}{\left( \sum_h n_{h21} n_{h1} / n_h \right)}
\]

It is always computed unless the denominator is zero. See Mantel and Haenszel (1959) and Agresti (2002) for more information.

To compute confidence limits for the common relative risk, PROC FREQ uses the Greenland and Robins (1985) variance estimate for \(\log(RR_{MH})\). The 100(1 − α/2)% confidence limits for the common relative risk are

\[
\left( RR_{MH} \times \exp(-z\hat{\theta}), \ RR_{MH} \times \exp(z\hat{\theta}) \right)
\]

where

\[
\hat{\theta}^2 = \text{Var}(\ln(RR_{MH})) = \frac{\sum_h (n_{h1} \cdot n_{h2} \cdot n_{h1} - n_{h11} n_{h21} n_h)^2}{\left( \sum_h n_{h11} n_{h2} / n_h \right) \left( \sum_h n_{h21} n_{h1} / n_h \right)}
\]

Logit Estimator The adjusted logit estimate of the common relative risk for column 1 is computed as

\[
RR_L = \exp \left( \sum_h w_h \ln(RR_h) / \sum w_h \right)
\]

and the corresponding 100(1 − α)% confidence limits are

\[
\left( RR_L \times \exp \left( -z / \sqrt{\sum_h w_h} \right), \ RR_L \times \exp \left( z / \sqrt{\sum_h w_h} \right) \right)
\]

where \(RR_h\) is the column 1 relative risk estimate for stratum \(h\) and

\[
w_h = 1 / \text{Var}(\ln(RR_h))
\]

If \(n_{h11}\) or \(n_{h21}\) is zero, then PROC FREQ adds 0.5 to each cell of the stratum before computing \(RR_h\) and \(w_h\) for the logit estimate. The procedure prints a warning when this occurs. See Kleinbaum, Kupper, and Morgenstern (1982, Sections 17.4 and 17.5) for details.

Breslow-Day Test for Homogeneity of the Odds Ratios

When you specify the CMH option, PROC FREQ computes the Breslow-Day test for stratified 2 × 2 tables. It tests the null hypothesis that the odds ratios for the \(q\) strata are equal. When the null hypothesis is true, the statistic has approximately a chi-square distribution with \(q–1\) degrees of freedom. See Breslow and Day (1980) and Agresti (2007) for more information.

The Breslow-Day statistic is computed as

\[
Q_{BD} = \sum_h (n_{h11} - E(n_{h11} | OR_{MH}))^2 / \text{Var}(n_{h11} | OR_{MH})
\]
where E and Var denote expected value and variance, respectively. The summation does not include any table with a zero row or column. If OR\textsubscript{MH} equals zero or if it is undefined, then PROC FREQ does not compute the statistic and prints a warning message.

For the Breslow-Day test to be valid, the sample size should be relatively large in each stratum, and at least 80% of the expected cell counts should be greater than 5. Note that this is a stricter sample size requirement than the requirement for the Cochran-Mantel-Haenszel test for $q \times 2 \times 2$ tables, in that each stratum sample size (not just the overall sample size) must be relatively large. Even when the Breslow-Day test is valid, it might not be very powerful against certain alternatives, as discussed in Breslow and Day (1980).

If you specify the BDT option, PROC FREQ computes the Breslow-Day test with Tarone’s adjustment, which subtracts an adjustment factor from $Q_{BD}$ to make the resulting statistic asymptotically chi-square. The Breslow-Day-Tarone statistic is computed as

$$Q_{BDT} = Q_{BD} - \left( \sum_h (n_{h11} - E(n_{h11} \mid OR_{MH})) \right)^2 / \sum_h \text{Var}(n_{h11} \mid OR_{MH})$$

See Tarone (1985); Jones et al. (1989); Breslow (1996) for more information.

### Zelen’s Exact Test for Equal Odds Ratios

If you specify the EQOR option in the EXACT statement, PROC FREQ computes Zelen’s exact test for equal odds ratios for stratified $2 \times 2$ tables. Zelen’s test is an exact counterpart to the Breslow-Day asymptotic test for equal odds ratios. The reference set for Zelen’s test includes all possible $q \times 2 \times 2$ tables with the same row, column, and stratum totals as the observed multiway table and with the same sum of cell (1,1) frequencies as the observed table. The test statistic is the probability of the observed $q \times 2 \times 2$ table conditional on the fixed margins, which is a product of hypergeometric probabilities.

The $p$-value for Zelen’s test is the sum of all table probabilities that are less than or equal to the observed table probability, where the sum is computed over all tables in the reference set determined by the fixed margins and the observed sum of cell (1,1) frequencies. This test is similar to Fisher’s exact test for two-way tables. For more information, see Zelen (1971); Hirji (2006); Agresti (1992). PROC FREQ computes Zelen’s exact test by using the polynomial multiplication algorithm of Hirji et al. (1996).

### Exact Confidence Limits for the Common Odds Ratio

If you specify the COMOR option in the EXACT statement, PROC FREQ computes exact confidence limits for the common odds ratio for stratified $2 \times 2$ tables. This computation assumes that the odds ratio is constant over all the $2 \times 2$ tables. Exact confidence limits are constructed from the distribution of $S = \sum_h n_{h11}$, conditional on the marginal totals of the $2 \times 2$ tables.

Because this is a discrete problem, the confidence coefficient for these exact confidence limits is not exactly $(1 - \alpha)$ but is at least $(1 - \alpha)$. Thus, these confidence limits are conservative. See Agresti (1992) for more information.


Conditional on the marginal totals of $2 \times 2$ table $h$, let the random variable $S_h$ denote the frequency of table cell (1,1). Given the row totals $n_{h1}$ and $n_{h2}$, and column totals $n_{h1}$ and $n_{h2}$, the lower and upper bounds for $S_h$ are $l_h$ and $u_h$,

$$l_h = \max (0, n_{h1} - n_{h2})$$

$$u_h = \min (n_{h1}, n_{h2})$$
Let $C_{sh}$ denote the hypergeometric coefficient,

$$C_{sh} = \binom{n_{h,1}}{s_h} \binom{n_{h,2}}{n_{h,1} - s_h}$$

and let $\phi$ denote the common odds ratio. Then the conditional distribution of $S_h$ is

$$P(S_h = s_h | n_{1.., n_{1.., n_{2..}}}) = C_{sh} \phi^{s_h} \sum_{x = l}^{u_h} C_x \phi^x$$

Summing over all the $2 \times 2$ tables, $S = \sum_h S_h$, and the lower and upper bounds of $S$ are $l$ and $u$,

$$l = \sum_h l_h \quad \text{and} \quad u = \sum_h u_h$$

The conditional distribution of the sum $S$ is

$$P(S = s | n_{1.., n_{1.., n_{2..}}}: h = 1, \ldots, q) = C_s \phi^s \sum_{x = l}^{u} C_x \phi^x$$

where

$$C_s = \sum_{s_1 + \ldots + s_q = s} \left( \prod_h C_{sh} \right)$$

Let $s_0$ denote the observed sum of cell (1,1) frequencies over the $q$ tables. The following two equations are solved iteratively for lower and upper confidence limits for the common odds ratio, $\phi_1$ and $\phi_2$:

$$\sum_{x = s_0}^{x = u} C_x \phi_1^x / \sum_{x = l}^{x = u} C_x \phi_1^x = \alpha/2$$

$$\sum_{x = l}^{x = u} C_x \phi_2^x / \sum_{x = l}^{x = u} C_x \phi_2^x = \alpha/2$$

When the observed sum $s_0$ equals the lower bound $l$, PROC FREQ sets the lower confidence limit to zero and determines the upper limit with level $\alpha$. Similarly, when the observed sum $s_0$ equals the upper bound $u$, PROC FREQ sets the upper confidence limit to infinity and determines the lower limit with level $\alpha$.

When you specify the COMOR option in the EXACT statement, PROC FREQ also computes the exact test that the common odds ratio equals one. Setting $\phi = 1$, the conditional distribution of the sum $S$ under the null hypothesis becomes

$$P_0(S = s | n_{1.., n_{1.., n_{2..}}}: h = 1, \ldots, q) = C_s / \sum_{x = l}^{x = u} C_x$$
The point probability for this exact test is the probability of the observed sum $s_0$ under the null hypothesis, conditional on the marginals of the stratified $2 \times 2$ tables, and is denoted by $P_0(s_0)$. The expected value of $S$ under the null hypothesis is

$$E_0(S) = \sum_{x = l}^{x = u} C_x$$

The one-sided exact $p$-value is computed from the conditional distribution as $P_0(S \geq s_0)$ or $P_0(S \leq s_0)$, depending on whether the observed sum $s_0$ is greater or less than $E_0(S)$,

$$P_1 = \begin{cases} P_0(S \geq s_0) = \sum_{x = s_0}^{x = u} C_x / \sum_{x = l}^{x = u} C_x & \text{if } s_0 > E_0(S) \\ P_0(S \leq s_0) = \sum_{x = l}^{x = s_0} C_x / \sum_{x = l}^{x = u} C_x & \text{if } s_0 \leq E_0(S) \end{cases}$$

PROC FREQ computes two-sided $p$-values for this test according to three different definitions. A two-sided $p$-value is computed as twice the one-sided $p$-value, setting the result equal to one if it exceeds one,

$$P_2^a = 2 \times P_1$$

Additionally, a two-sided $p$-value is computed as the sum of all probabilities less than or equal to the point probability of the observed sum $s_0$, summing over all possible values of $s$, $l \leq s \leq u$,

$$P_2^b = \sum_{l \leq s \leq u: P_0(s) \leq P_0(s_0)} P_0(s)$$

Also, a two-sided $p$-value is computed as the sum of the one-sided $p$-value and the corresponding area in the opposite tail of the distribution, equidistant from the expected value,

$$P_2^c = P_0(|S - E_0(S)| \geq |s_0 - E_0(S)|)$$

**Gail-Simon Test for Qualitative Interactions**

The GAILSIMON option in the TABLES statement provides the Gail-Simon test for qualitative interaction for stratified $2 \times 2$ tables. For details, see Gail and Simon (1985); Silvapulle (2001); Dmitrienko et al. (2005).

The Gail-Simon test is based on the risk differences in stratified $2 \times 2$ tables, where the risk difference is defined as the row 1 risk (proportion in column 1) minus the row 2 risk. See the section “Risks and Risk Differences” on page 174 for details. By default, the procedure uses column 1 risks to compute the Gail-Simon test. If you specify the GAILSIMON(COLUMN=2) option, the procedure uses column 2 risks.

PROC FREQ computes the Gail-Simon test statistics as described in Gail and Simon (1985),

$$Q^- = \sum_h (d_h/s_h)^2 I(d_h > 0)$$

$$Q^+ = \sum_h (d_h/s_h)^2 I(d_h < 0)$$

$$Q = \min (Q^-, Q^+)$$
where \( d_h \) is the risk difference in table \( h \), \( s_h \) is the standard error of the risk difference, and \( I(d_h > 0) \) equals 1 if \( d_h > 0 \) and 0 otherwise. Similarly, \( I(d_h < 0) \) equals 1 if \( d_h < 0 \) and 0 otherwise. The \( q \times 2 \times 2 \) tables (strata) are indexed by \( h = 1, 2, \ldots, q \).

The \( p \)-values for the Gail-Simon statistics are computed as

\[
\begin{align*}
p(Q^-) &= \sum_h (1 - F_h(Q^-)) B(h; n = q, p = 0.5) \\
p(Q^+) &= \sum_h (1 - F_h(Q^+)) B(h; n = q, p = 0.5) \\
p(Q) &= \sum_{h=1}^{q-1} (1 - F_h(Q)) B(h; n = (q-1), p = 0.5)
\end{align*}
\]

where \( F_h(\cdot) \) is the cumulative chi-square distribution function with \( h \) degrees of freedom and \( B(h; n, p) \) is the binomial probability function with parameters \( n \) and \( p \). The statistic \( Q \) tests the null hypothesis of no qualitative interaction. The statistic \( Q^- \) tests the null hypothesis of positive risk differences. A small \( p \)-value for \( Q^- \) indicates negative differences; similarly, a small \( p \)-value for \( Q^+ \) indicates positive risk differences.

### Exact Statistics

Exact statistics can be useful in situations where the asymptotic assumptions are not met, and so the asymptotic \( p \)-values are not close approximations for the true \( p \)-values. Standard asymptotic methods involve the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. When the sample size is not large, asymptotic results might not be valid, with the asymptotic \( p \)-values differing perhaps substantially from the exact \( p \)-values. Asymptotic results might also be unreliable when the distribution of the data is sparse, skewed, or heavily tied. See Agresti (2007) and Bishop, Fienberg, and Holland (1975) for more information. Exact computations are based on the statistical theory of exact conditional inference for contingency tables, reviewed by Agresti (1992).

In addition to computation of exact \( p \)-values, PROC FREQ provides the option of estimating exact \( p \)-values by Monte Carlo simulation. This can be useful for problems that are so large that exact computations require a great amount of time and memory, but for which asymptotic approximations might not be sufficient.

Exact statistics are available for many PROC FREQ tests. For one-way tables, PROC FREQ provides exact \( p \)-values for the binomial proportion tests and the chi-square goodness-of-fit test. Exact (Clopper-Pearson) confidence limits are available for the binomial proportion. For two-way tables, PROC FREQ provides exact \( p \)-values for the following tests: Pearson chi-square test, likelihood ratio chi-square test, Mantel-Haenszel chi-square test, Fisher’s exact test, Jonckheere-Terpstra test, and Cochran-Armitage test for trend. PROC FREQ also computes exact \( p \)-values for tests of the following statistics: Kendall’s tau-\( b \), Stuart’s tau-\( c \), Somers’ \( D(C|R) \), Somers’ \( D(R|C) \), Pearson correlation coefficient, Spearman correlation coefficient, simple kappa coefficient, and weighted kappa coefficient. For \( 2 \times 2 \) tables, PROC FREQ provides McNemar’s exact test and exact confidence limits for the odds ratio. PROC FREQ also provides exact unconditional confidence limits for the proportion (risk) difference and for the relative risk. For stratified \( 2 \times 2 \) tables, PROC FREQ provides Zelen’s exact test for equal odds ratios, exact confidence limits for the common odds ratio, and an exact test for the common odds ratio.

The following sections summarize the exact computational algorithms, define the exact \( p \)-values that PROC FREQ computes, discuss the computational resource requirements, and describe the Monte Carlo estimation option.
**Computational Algorithms**

PROC FREQ computes exact $p$-values for general $R \times C$ tables by using the network algorithm developed by Mehta and Patel (1983). This algorithm provides a substantial advantage over direct enumeration, which can be very time-consuming and feasible only for small problems. See Agresti (1992) for a review of algorithms for computation of exact $p$-values, and see Mehta, Patel, and Tsiatis (1984) and Mehta, Patel, and Senchaudhuri (1991) for information about the performance of the network algorithm.

The reference set for a given contingency table is the set of all contingency tables with the observed marginal row and column sums. Corresponding to this reference set, the network algorithm forms a directed acyclic network consisting of nodes in a number of stages. A path through the network corresponds to a distinct table in the reference set. The distances between nodes are defined so that the total distance of a path through the network is the corresponding value of the test statistic. At each node, the algorithm computes the shortest and longest path distances for all the paths that pass through that node. For statistics that can be expressed as a linear combination of cell frequencies multiplied by increasing row and column scores, PROC FREQ computes shortest and longest path distances by using the algorithm of Agresti, Mehta, and Patel (1990). For statistics of other forms, PROC FREQ computes an upper bound for the longest path and a lower bound for the shortest path by following the approach of Valz and Thompson (1994).

The longest and shortest path distances or bounds for a node are compared to the value of the test statistic to determine whether all paths through the node contribute to the $p$-value, none of the paths through the node contribute to the $p$-value, or neither of these situations occurs. If all paths through the node contribute, the $p$-value is incremented accordingly, and these paths are eliminated from further analysis. If no paths contribute, these paths are eliminated from the analysis. Otherwise, the algorithm continues, still processing this node and the associated paths. The algorithm finishes when all nodes have been accounted for.

In applying the network algorithm, PROC FREQ uses full numerical precision to represent all statistics, row and column scores, and other quantities involved in the computations. Although it is possible to use rounding to improve the speed and memory requirements of the algorithm, PROC FREQ does not do this because it can result in reduced accuracy of the $p$-values.

For one-way tables, PROC FREQ computes the exact chi-square goodness-of-fit test by the method of Radlow and Alf (1975). PROC FREQ generates all possible one-way tables with the observed total sample size and number of categories. For each possible table, PROC FREQ compares its chi-square value with the value for the observed table. If the table’s chi-square value is greater than or equal to the observed chi-square, PROC FREQ increments the exact $p$-value by the probability of that table, which is calculated under the null hypothesis by using the multinomial frequency distribution. By default, the null hypothesis states that all categories have equal proportions. If you specify null hypothesis proportions or frequencies by using the TESTP= or TESTF= option in the TABLES statement, then PROC FREQ calculates the exact chi-square test based on that null hypothesis.

Other exact computations are described in sections about the individual statistics. See the section “Binomial Proportion” on page 168 for details about how PROC FREQ computes exact confidence limits and tests for the binomial proportion. See the section “Odds Ratio and Relative Risks for 2 x 2 Tables” on page 186 for information about computation of exact confidence limits for the odds ratio for $2 \times 2$ tables. Also, see the sections “Exact Unconditional Confidence Limits for the Risk Difference” on page 183, “Exact Confidence Limits for the Common Odds Ratio” on page 203, and “Zelen’s Exact Test for Equal Odds Ratios” on page 203.

**Definition of $p$-Values**

For several tests in PROC FREQ, the test statistic is nonnegative, and large values of the test statistic indicate a departure from the null hypothesis. Such nondirectional tests include the Pearson chi-square, the likelihood ratio chi-square, the Mantel-Haenszel chi-square, Fisher’s exact test for tables larger than $2 \times 2$, McNemar’s
test, and the one-way chi-square goodness-of-fit test. The exact $p$-value for a nondirectional test is the sum of probabilities for those tables having a test statistic greater than or equal to the value of the observed test statistic.

There are other tests where it might be appropriate to test against either a one-sided or a two-sided alternative hypothesis. For example, when you test the null hypothesis that the true parameter value equals 0 ($T = 0$), the alternative of interest might be one-sided ($T ≤ 0$, or $T ≥ 0$) or two-sided ($T ≠ 0$). Such tests include the Pearson correlation coefficient, Spearman correlation coefficient, Jonckheere-Terpstra test, Cochran-Armitage test for trend, simple kappa coefficient, and weighted kappa coefficient. For these tests, PROC FREQ displays the right-sided $p$-value when the observed value of the test statistic is greater than its expected value. The right-sided $p$-value is the sum of probabilities for those tables for which the test statistic is greater than or equal to the observed test statistic. Otherwise, when the observed test statistic is less than or equal to the expected value, PROC FREQ displays the left-sided $p$-value. The left-sided $p$-value is the sum of probabilities for those tables for which the test statistic is less than or equal to the one observed. The one-sided $p$-value $P_1$ can be expressed as

$$P_1 = \begin{cases} \text{Prob( Test Statistic ≥ } t \text{ ) if } t > E_0(T) \\ \text{Prob( Test Statistic ≤ } t \text{ ) if } t ≤ E_0(T) \end{cases}$$

where $t$ is the observed value of the test statistic and $E_0(T)$ is the expected value of the test statistic under the null hypothesis. PROC FREQ computes the two-sided $p$-value as the sum of the one-sided $p$-value and the corresponding area in the opposite tail of the distribution of the statistic, equidistant from the expected value. The two-sided $p$-value $P_2$ can be expressed as

$$P_2 = \text{Prob( |Test Statistic − E_0(T)| ≥ |t − E_0(T)| )}$$

If you specify the POINT option in the EXACT statement, PROC FREQ provides exact point probabilities for the exact tests. The exact point probability is the exact probability that the test statistic equals the observed value.

If you specify the MIDP option in the EXACT statement, PROC FREQ provides exact mid-$p$-values. The exact mid $p$-value is defined as the exact $p$-value minus half the exact point probability, which equals the average of $\text{Prob( Test Statistic ≥ } t \text{ )}$ and $\text{Prob( Test Statistic > } t \text{ )}$ for a right-sided test. The exact mid $p$-value is smaller and less conservative than the non-adjusted exact $p$-value. For more information, see Agresti (2013, section 1.1.4) and Hirji (2006, sections 2.5 and 2.11.1).

**Computational Resources**

PROC FREQ uses relatively fast and efficient algorithms for exact computations. These recently developed algorithms, together with improvements in computer power, now make it feasible to perform exact computations for data sets where previously only asymptotic methods could be applied. Nevertheless, there are still large problems that might require a prohibitive amount of time and memory for exact computations, depending on the speed and memory available on your computer. For large problems, consider whether exact methods are really needed or whether asymptotic methods might give results quite close to the exact results, while requiring much less computer time and memory. When asymptotic methods might not be sufficient for such large problems, consider using Monte Carlo estimation of exact $p$-values, as described in the section “Monte Carlo Estimation” on page 209.

A formula does not exist that can predict in advance how much time and memory are needed to compute an exact $p$-value for a certain problem. The time and memory required depend on several factors, including which test is being performed, the total sample size, the number of rows and columns, and the specific arrangement of the observations into table cells. Generally, larger problems (in terms of total sample size,
number of rows, and number of columns) tend to require more time and memory. Additionally, for a fixed total sample size, time and memory requirements tend to increase as the number of rows and columns increases, because this corresponds to an increase in the number of tables in the reference set. Also for a fixed sample size, time and memory requirements increase as the marginal row and column totals become more homogeneous. See Agresti, Mehta, and Patel (1990) and Gail and Mantel (1977) for more information.

At any time while PROC FREQ is computing exact p-values, you can terminate the computations by pressing the system interrupt key sequence (see the SAS Companion for your system) and choosing to stop computations. After you terminate exact computations, PROC FREQ completes all other remaining tasks. The procedure produces the requested output and reports missing values for any exact p-values that were not computed by the time of termination.

You can also use the MAXTIME= option in the EXACT statement to limit the amount of time PROC FREQ uses for exact computations. You specify a MAXTIME= value that is the maximum amount of clock time (in seconds) that PROC FREQ can use to compute an exact p-value. If PROC FREQ does not finish computing an exact p-value within that time, it terminates the computation and completes all other remaining tasks.

**Monte Carlo Estimation**

If you specify the option MC in the EXACT statement, PROC FREQ computes Monte Carlo estimates of the exact p-values instead of directly computing the exact p-values. Monte Carlo estimation can be useful for large problems that require a great amount of time and memory for exact computations but for which asymptotic approximations might not be sufficient. To describe the precision of each Monte Carlo estimate, PROC FREQ provides the asymptotic standard error and 100(1 − α)% confidence limits. The confidence level α is determined by the ALPHA= option in the EXACT statement, which, by default, equals 0.01 and produces 99% confidence limits. The N= option in the EXACT statement specifies the number of samples that PROC FREQ uses for Monte Carlo estimation; the default is 10000 samples. You can specify a larger value for n to improve the precision of the Monte Carlo estimates. Because larger values of n generate more samples, the computation time increases. Alternatively, you can specify a smaller value of n to reduce the computation time.

To compute a Monte Carlo estimate of an exact p-value, PROC FREQ generates a random sample of tables with the same total sample size, row totals, and column totals as the observed table. PROC FREQ uses the algorithm of Agresti, Wackerly, and Boyett (1979), which generates tables in proportion to their hypergeometric probabilities conditional on the marginal frequencies. For each sample table, PROC FREQ computes the value of the test statistic and compares it to the value for the observed table. When estimating a right-sided p-value, PROC FREQ counts all sample tables for which the test statistic is greater than or equal to the observed test statistic. Then the p-value estimate equals the number of these tables divided by the total number of tables sampled.

\[ \hat{P}_{MC} = \frac{M}{N} \]

- \( M \) = number of samples with (Test Statistic \( \geq t \))
- \( N \) = total number of samples
- \( t \) = observed Test Statistic

PROC FREQ computes left-sided and two-sided p-value estimates in a similar manner. For left-sided p-values, PROC FREQ evaluates whether the test statistic for each sampled table is less than or equal to the observed test statistic. For two-sided p-values, PROC FREQ examines the sample test statistics according to the expression for \( P_2 \) given in the section “Definition of p-Values” on page 207.
The variable $M$ is a binomially distributed variable with $N$ trials and success probability $p$. It follows that the asymptotic standard error of the Monte Carlo estimate is

$$se(\hat{P}_{MC}) = \sqrt{\hat{P}_{MC} (1 - \hat{P}_{MC}) / (N - 1)}$$

PROC FREQ constructs asymptotic confidence limits for the $p$-values according to

$$\hat{P}_{MC} \pm \left( z_{\alpha/2} \times se(\hat{P}_{MC}) \right)$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution and the confidence level $\alpha$ is determined by the ALPHA= option in the EXACT statement.

When the Monte Carlo estimate $\hat{P}_{MC}$ equals 0, PROC FREQ computes the confidence limits for the $p$-value as

$$( 0, \ 1 - \alpha^{(1/N)} )$$

When the Monte Carlo estimate $\hat{P}_{MC}$ equals 1, PROC FREQ computes the confidence limits as

$$( \alpha^{(1/N)}, \ 1 )$$

---

**Computational Resources**

For each variable in a table request, PROC FREQ stores all of the levels in memory. If all variables are numeric and not formatted, this requires about 84 bytes for each variable level. When there are character variables or formatted numeric variables, the memory that is required depends on the formatted variable lengths, with longer formatted lengths requiring more memory. The number of levels for each variable is limited only by the largest integer that your operating environment can store.

For any single crosstabulation table requested, PROC FREQ builds the entire table in memory, regardless of whether the table has zero cell counts. Thus, if the numeric variables $A$, $B$, and $C$ each have 10 levels, PROC FREQ requires 2520 bytes to store the variable levels for the table request $A*B*C$, as follows:

$$3 \text{ variables} \times 10 \text{ levels/variable} \times 84 \text{ bytes/level}$$

In addition, PROC FREQ requires 8000 bytes to store the table cell frequencies

$$1000 \text{ cells} \times 8 \text{ bytes/cell}$$

even though there might be only 10 observations.

When the variables have many levels or when there are many multiway tables, your computer might not have enough memory to construct the tables. If PROC FREQ runs out of memory while constructing tables, it stops collecting levels for the variable with the most levels and returns the memory that is used by that variable. The procedure then builds the tables that do not contain the disabled variables.

If there is not enough memory for your table request and if increasing the available memory is impractical, you can reduce the number of multiway tables or variable levels. If you are not using the CMH or AGREE
option in the TABLES statement to compute statistics across strata, reduce the number of multiway tables by using PROC SORT to sort the data set by one or more of the variables or by using the DATA step to create an index for the variables. Then remove the sorted or indexed variables from the TABLES statement and include a BY statement that uses these variables. You can also reduce memory requirements by using a FORMAT statement in the PROC FREQ step to reduce the number of levels. Additionally, reducing the formatted variable lengths reduces the amount of memory that is needed to store the variable levels. For more information about using formats, see the section “Grouping with Formats” on page 147.

Output Data Sets

PROC FREQ produces two types of output data sets that you can use with other statistical and reporting procedures. You can request these data sets as follows:

- Specify the OUT= option in a TABLES statement. This creates an output data set that contains frequency or crosstabulation table counts and percentages
- Specify an OUTPUT statement. This creates an output data set that contains statistics.

PROC FREQ does not display the output data sets. Use PROC PRINT, PROC REPORT, or any other SAS reporting tool to display an output data set.

In addition to these two output data sets, you can create a SAS data set from any piece of PROC FREQ output by using the Output Delivery System. See the section “ODS Table Names” on page 222 for more information.

Contents of the TABLES Statement Output Data Set

The OUT= option in the TABLES statement creates an output data set that contains one observation for each combination of variable values (or table cell) in the last table request. By default, each observation contains the frequency and percentage for the table cell. When the input data set contains missing values, the output data set also contains an observation with the frequency of missing values. The output data set includes the following variables:

- BY variables
- table request variables, such as A, B, C, and D in the table request A*B*C*D
- COUNT, which contains the table cell frequency
- PERCENT, which contains the table cell percentage

If you specify the OUTEXPECT option in the TABLES statement for a two-way or multiway table, the output data set also includes expected frequencies. If you specify the OUTPCT option for a two-way or multiway table, the output data set also includes row, column, and table percentages. The additional variables are as follows:

- EXPECTED, which contains the expected frequency
- PCT_TABL, which contains the percentage of two-way table frequency, for n-way tables where n > 2
Chapter 3: The FREQ Procedure

- **PCT_ROW**, which contains the percentage of row frequency
- **PCT_COL**, which contains the percentage of column frequency

If you specify the OUTCUM option in the TABLES statement for a one-way table, the output data set also includes cumulative frequencies and cumulative percentages. The additional variables are as follows:

- **CUM_FREQ**, which contains the cumulative frequency
- **CUM_PCT**, which contains the cumulative percentage

The OUTCUM option has no effect for two-way or multiway tables.

The following PROC FREQ statements create an output data set of frequencies and percentages:

```sas
proc freq;
    tables A A*B / out=D;
run;
```

The output data set `D` contains frequencies and percentages for the table of `A` by `B`, which is the last table request listed in the TABLES statement. If `A` has two levels (1 and 2), `B` has three levels (1, 2, and 3), and no table cell count is zero or missing, then the output data set `D` includes six observations, one for each combination of `A` and `B` levels. The first observation corresponds to `A=1` and `B=1`; the second observation corresponds to `A=1` and `B=2`; and so on. The data set includes the variables `COUNT` and `PERCENT`. The value of `COUNT` is the number of observations with the given combination of `A` and `B` levels. The value of `PERCENT` is the percentage of the total number of observations with that `A` and `B` combination.

When PROC FREQ combines different variable values into the same formatted level, the output data set contains the smallest internal value for the formatted level. For example, suppose a variable `X` has the values 1.1, 1.4, 1.7, 2.1, and 2.3. When you submit the statement

```sas
format X 1.;
```

in a PROC FREQ step, the formatted levels listed in the frequency table for `X` are 1 and 2. If you create an output data set with the frequency counts, the internal values of the levels of `X` are 1.1 and 1.7. To report the internal values of `X` when you display the output data set, use a format of 3.1 for `X`.

### Contents of the OUTPUT Statement Output Data Set

The **OUTPUT** statement creates a SAS data set that contains statistics computed by PROC FREQ. Table 3.7 lists the statistics that can be stored in the output data set. You identify which statistics to include by specifying **output-options**. See the description of the OUTPUT statement for details.

If you specify multiple **TABLES** statements or multiple table requests in a single **TABLES** statement, the contents of the output data set correspond to the last table request.

For a one-way table or a two-way table, the output data set contains one observation that stores the requested statistics for the table. For a multiway table, the output data set contains an observation for each two-way table (stratum) of the multiway crosstabulation. If you request summary statistics for the multiway table, the output data set also contains an observation that stores the across-strata summary statistics. If you use a **BY**
statement, the output data set contains an observation (for one-way or two-way tables) or set of observations (for multiway tables) for each BY group.

The OUTPUT data set can include the following variables:

- BY variables
- Variables that identify the stratum for multiway tables, such as A and B in the table request $A*B*C*D$
- Variables that contain the specified statistics

In addition to the specified estimate or test statistic, the output data set includes associated values such as standard errors, confidence limits, $p$-values, and degrees of freedom.

PROC FREQ constructs variable names for the statistics in the output data set by enclosing the *output-option* names in underscores. Variable names for the corresponding standard errors, confidence limits, $p$-values, and degrees of freedom are formed by combining the *output-option* names with prefixes that identify the associated values. Table 3.20 lists the prefixes and their descriptions.

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E_</td>
<td>Asymptotic standard error (ASE)</td>
</tr>
<tr>
<td>L_</td>
<td>Lower confidence limit</td>
</tr>
<tr>
<td>U_</td>
<td>Upper confidence limit</td>
</tr>
<tr>
<td>E0_</td>
<td>Null hypothesis ASE</td>
</tr>
<tr>
<td>Z_</td>
<td>Standardized value</td>
</tr>
<tr>
<td>DF_</td>
<td>Degrees of freedom</td>
</tr>
<tr>
<td>P_</td>
<td>$p$-value</td>
</tr>
<tr>
<td>P2_</td>
<td>Two-sided $p$-value</td>
</tr>
<tr>
<td>PL_</td>
<td>Left-sided $p$-value</td>
</tr>
<tr>
<td>PR_</td>
<td>Right-sided $p$-value</td>
</tr>
<tr>
<td>XP_</td>
<td>Exact $p$-value</td>
</tr>
<tr>
<td>XP2_</td>
<td>Exact two-sided $p$-value</td>
</tr>
<tr>
<td>XPL_</td>
<td>Exact left-sided $p$-value</td>
</tr>
<tr>
<td>XPR_</td>
<td>Exact right-sided $p$-value</td>
</tr>
<tr>
<td>XPT_</td>
<td>Exact point probability</td>
</tr>
<tr>
<td>XMP_</td>
<td>Exact mid $p$-value</td>
</tr>
<tr>
<td>XL_</td>
<td>Exact lower confidence limit</td>
</tr>
<tr>
<td>XU_</td>
<td>Exact upper confidence limit</td>
</tr>
</tbody>
</table>

For example, the *PCHI output-option* in the OUTPUT statement includes the Pearson chi-square test in the output data set. The variable names for the Pearson chi-square statistic, its degrees of freedom, and the corresponding $p$-value are _PCHI_, DF_PCHI, and P_PCHI, respectively. For variables that were added to the output data set before SAS/STAT 8.2, PROC FREQ truncates the variable name to eight characters when the length of the prefix plus the *output-option* name exceeds eight characters.
Displayed Output

Number of Variable Levels Table

If you specify the NLEVELS option in the PROC FREQ statement, PROC FREQ displays the “Number of Variable Levels” table. This table provides the number of levels for all variables named in the TABLES statements. PROC FREQ determines the variable levels from the formatted variable values. See “Grouping with Formats” on page 147 for details. The “Number of Variable Levels” table contains the following information:

- Variable name
- Levels, which is the total number of levels of the variable
- Number of Nonmissing Levels, if there are missing levels for any of the variables
- Number of Missing Levels, if there are missing levels for any of the variables

One-Way Frequency Tables

PROC FREQ displays one-way frequency tables for all one-way table requests in the TABLES statements, unless you specify the NOPRINT option in the PROC FREQ statement or the NOPRINT option in the TABLES statement. For a one-way table showing the frequency distribution of a single variable, PROC FREQ displays the name of the variable and its values. For each variable value or level, PROC FREQ displays the following information:

- Frequency count, which is the number of observations in the level
- Test Frequency count, if you specify the CHISQ and TESTF= options to request a chi-square goodness-of-fit test for specified frequencies
- Percent, which is the percentage of the total number of observations. (The NOPERCENT option suppresses this information.)
- Test Percent, if you specify the CHISQ and TESTP= options to request a chi-square goodness-of-fit test for specified percents. (The NOPERCENT option suppresses this information.)
- Cumulative Frequency count, which is the sum of the frequency counts for that level and all other levels listed above it in the table. The last cumulative frequency is the total number of nonmissing observations. (The NOCUM option suppresses this information.)
- Cumulative Percent, which is the percentage of the total number of observations in that level and in all other levels listed above it in the table. (The NOCUM or the NOPERCENT option suppresses this information.)

The one-way table also displays the Frequency Missing, which is the number of observations with missing values.
Statistics for One-Way Frequency Tables

For one-way tables, two statistical options are available in the TABLES statement. The CHISQ option provides a chi-square goodness-of-fit test, and the BINOMIAL option provides binomial proportion statistics and tests. PROC FREQ displays the following information, unless you specify the NOPRINT option in the PROC FREQ statement:

- If you specify the CHISQ option for a one-way table, PROC FREQ provides a chi-square goodness-of-fit test, displaying the Chi-Square statistic, the degrees of freedom (DF), and the probability value (Pr > ChiSq). If you specify the CHISQ option in the EXACT statement, PROC FREQ also displays the exact probability value for this test. If you specify the POINT option with the CHISQ option in the EXACT statement, PROC FREQ displays the exact point probability for the test statistic. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact mid p-value for the chi-square test.

- If you specify the BINOMIAL option for a one-way table, PROC FREQ displays the estimate of the binomial Proportion, which is the proportion of observations in the first class listed in the one-way table. PROC FREQ also displays the asymptotic standard error (ASE) and the asymptotic (Wald) and exact (Clopper-Pearson) confidence limits by default. For the binomial proportion test, PROC FREQ displays the asymptotic standard error under the null hypothesis (ASE Under H0), the standardized test statistic (Z), and the one-sided and two-sided probability values.

If you specify the BINOMIAL option in the EXACT statement, PROC FREQ also displays the exact one-sided and two-sided probability values for this test. If you specify the POINT option with the BINOMIAL option in the EXACT statement, PROC FREQ displays the exact point probability for the test. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact mid p-value for the binomial proportion test.

- If you request additional binomial confidence limits by specifying binomial-options, PROC FREQ provides a table that displays the lower and upper confidence limits for each type that you request. In addition to the Wald and exact (Clopper-Pearson) confidence limits, you can request Agresti-Coull, Jeffreys, and Wilson (score) confidence limits for the binomial proportion.

- If you request a binomial noninferiority or superiority test by specifying the NONINF or SUP binomial-option, PROC FREQ displays a Noninferiority Analysis or Superiority Analysis table that contains the following information: the binomial Proportion, the test ASE (under H0 or Sample), the test statistic Z, the probability value, the noninferiority or superiority limit, and the test confidence limits. If you specify the BINOMIAL option in the EXACT statement, PROC FREQ also provides the exact probability value for the test, and exact test confidence limits.

- If you request a binomial equivalence test by specifying the EQUIV binomial-option, PROC FREQ displays an Equivalence Analysis table that contains the following information: binomial Proportion and the test ASE (under H0 or Sample). PROC FREQ displays two one-sided tests (TOST) for equivalence, which include test statistics (Z) and probability values for the Lower and Upper tests, together with the Overall probability value. PROC FREQ also displays the equivalence limits and the test-based confidence limits. If you specify the BINOMIAL option in the EXACT statement, PROC FREQ provides exact probability values for the TOST and exact test-based confidence limits.
Two-Way and Multiway Tables

PROC FREQ displays all multiway table requests in the TABLES statements, unless you specify the NOPRINT option in the PROC FREQ statement or the NOPRINT option in the TABLES statement.

For two-way to multiway crosstabulation tables, the values of the last variable in the table request form the table columns. The values of the next-to-last variable form the rows. Each level (or combination of levels) of the other variables forms one stratum.

There are three ways to display multiway tables in PROC FREQ. By default, PROC FREQ displays multiway tables as separate two-way crosstabulation tables for each stratum of the multiway table. Also by default, PROC FREQ displays these two-way crosstabulation tables in table cell format. Alternatively, if you specify the CROSSLIST option, PROC FREQ displays the two-way crosstabulation tables in ODS column format. If you specify the LIST option, PROC FREQ displays multiway tables in list format, which presents the entire multiway crosstabulation in a single table.

Crosstabulation Tables

By default, PROC FREQ displays two-way crosstabulation tables in table cell format. The row variable values are listed down the side of the table, the column variable values are listed across the top of the table, and each row and column variable level combination forms a table cell.

Each cell of a crosstabulation table can contain the following information:

- Frequency, which is the number of observations in the table cell. (The NOFREQ option suppresses this information.)
- Expected frequency under the hypothesis of independence, if you specify the EXPECTED option
- Deviation of the cell frequency from the expected value, if you specify the DEVIATION option
- Cell Chi-Square, which is the cell’s contribution to the total chi-square statistic, if you specify the CELLCHI2 option
- Tot Pct, which is the cell’s percentage of the total multiway table frequency, for n-way tables when \( n \geq 2 \), if you specify the TOTPCT option
- Percent, which is the cell’s percentage of the total (two-way table) frequency. (The NOPERCENT option suppresses this information.)
- Row Pct, or the row percentage, which is the cell’s percentage of the total frequency for its row. (The NOROW option suppresses this information.)
- Col Pct, or column percentage, which is the cell’s percentage of the total frequency for its column. (The NOCOL option suppresses this information.)
- Cumulative Col%, or cumulative column percentage, if you specify the CUMCOL option

The table also displays the Frequency Missing, which is the number of observations with missing values.
**CROSSLIST Tables**

If you specify the CROSSLIST option, PROC FREQ displays two-way crosstabulation tables in ODS column format. The CROSSLIST column format is different from the default crosstabulation table cell format, but the CROSSLIST table provides the same information (frequencies, percentages, and other statistics) as the default crosstabulation table.

In the CROSSLIST table format, the rows of the display correspond to the crosstabulation table cells, and the columns of the display correspond to descriptive statistics such as frequencies and percentages. Each table cell is identified by the values of its TABLES row and column variable levels, with all column variable levels listed within each row variable level. The CROSSLIST table also provides row totals, column totals, and overall table totals.

For a crosstabulation table in CROSSLIST format, PROC FREQ displays the following information:

- the row variable name and values
- the column variable name and values
- Frequency, which is the number of observations in the table cell. (The NOFREQ option suppresses this information.)
- Expected cell frequency under the hypothesis of independence, if you specify the EXPECTED option
- Deviation of the cell frequency from the expected value, if you specify the DEVIATION option
- Standardized Residual, if you specify the CROSSLIST(STDRES) option
- Pearson Residual, if you specify the CROSSLIST(PEARSONRES) option
- Cell Chi-Square, which is the cell’s contribution to the total chi-square statistic, if you specify the CELLCHI2 option
- Total Percent, which is the cell’s percentage of the total multiway table frequency, for \( n \)-way tables when \( n > 2 \), if you specify the TOTPCT option
- Percent, which is the cell’s percentage of the total (two-way table) frequency. (The NOPERCENT option suppresses this information.)
- Row Percent, which is the cell’s percentage of the total frequency for its row. (The NOROW option suppresses this information.)
- Column Percent, the cell’s percentage of the total frequency for its column. (The NOCOL option suppresses this information.)

The table also displays the Frequency Missing, which is the number of observations with missing values.

**LIST Tables**

If you specify the LIST option in the TABLES statement, PROC FREQ displays multiway tables in a list format rather than as crosstabulation tables. The LIST option displays the entire multiway table in one table, instead of displaying a separate two-way table for each stratum. The LIST option is not available when you also request statistical options. Unlike the default crosstabulation output, the LIST output does not
display row percentages, column percentages, and optional information such as expected frequencies and cell chi-squares.

For a multiway table in list format, PROC FREQ displays the following information:

- the variable names and values
- Frequency, which is the number of observations in the level (with the indicated variable values)
- Percent, which is the level’s percentage of the total number of observations. (The NOPERCENT option suppresses this information.)
- Cumulative Frequency, which is the accumulated frequency of the level and all other levels listed above it in the table. The last cumulative frequency in the table is the total number of nonmissing observations. (The NOCUM option suppresses this information.)
- Cumulative Percent, which is the accumulated percentage of the level and all other levels listed above it in the table. (The NOCUM or the NOPERCENT option suppresses this information.)

The table also displays the Frequency Missing, which is the number of observations with missing values.

Statistics for Two-Way and Multiway Tables

PROC FREQ computes statistical tests and measures for crosstabulation tables, depending on which statements and options you specify. You can suppress the display of these results by specifying the NOPRINT option in the PROC FREQ statement. With any of the following information, PROC FREQ also displays the Sample Size and the Frequency Missing.

- If you specify the SCOROUT option in the TABLES statement, PROC FREQ displays the Row Scores and Column Scores that it uses for statistical computations. The Row Scores table displays the row variable values and the Score corresponding to each value. The Column Scores table displays the column variable values and the corresponding Scores. PROC FREQ also identifies the score type used to compute the row and column scores. You can specify the score type with the SCORES= option in the TABLES statement.
- If you specify the CHISQ option, PROC FREQ displays the following statistics for each two-way table: Pearson Chi-Square, Likelihood Ratio Chi-Square, Continuity-Adjusted Chi-Square (for $2 \times 2$ tables), Mantel-Haenszel Chi-Square, the Phi Coefficient, the Contingency Coefficient, and Cramér’s V. For each test statistic, PROC FREQ also displays the degrees of freedom (DF) and the probability value (Prob).
- If you specify the CHISQ option for $2 \times 2$ tables, PROC FREQ also displays Fisher’s exact test. The test output includes the cell (1,1) frequency (F), the exact left-sided and right-sided probability values, the table probability (P), and the exact two-sided probability value. If you specify the POINT option in the EXACT statement, PROC FREQ displays the exact point probability for Fisher’s exact test. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the Mid $p$-Value for the test.
- If you specify the FISHER option in the TABLES statement (or, equivalently, the FISHER option in the EXACT statement), PROC FREQ displays Fisher’s exact test for tables larger than $2 \times 2$. The test output includes the table probability (P) and the probability value. If you specify the POINT option in the EXACT statement, PROC FREQ displays the exact point probability for Fisher’s exact test. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the Mid $p$-Value for the test.
• If you specify the PCHI, LRCHI, or MHCHI option in the EXACT statement, PROC FREQ displays the corresponding exact test: Pearson Chi-Square, Likelihood Ratio Chi-Square, or Mantel-Haenszel Chi-Square, respectively. The test output includes the test statistic, the degrees of freedom (DF), and the asymptotic and exact probability values. If you also specify the POINT option in the EXACT statement, PROC FREQ displays the point probability for each exact test requested. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact mid-p-value for each test. If you specify the CHISQ option in the EXACT statement, PROC FREQ displays exact probability values for all three of these chi-square tests.

• If you specify the MEASURES option, PROC FREQ displays the following statistics and their asymptotic standard errors (ASE) for each two-way table: Gamma, Kendall’s Tau-b, Stuart’s Tau-c, Somers’ D(C|R), Somers’ D(R|C), Pearson Correlation, Spearman Correlation, Lambda Asymmetric (C|R), Lambda Asymmetric (R|C), Lambda Symmetric, Uncertainty Coefficient (C|R), Uncertainty Coefficient (R|C), and Uncertainty Coefficient Symmetric. If you specify the CL option, PROC FREQ also displays confidence limits for these measures.

• If you specify the PLCORR option, PROC FREQ displays the tetrachoric correlation for 2 × 2 tables or the polychoric correlation for larger tables. In addition, PROC FREQ displays the MEASURES output listed earlier, even if you do not also specify the MEASURES option.

• If you specify the GAMMA, KENTB, STUTC, SMDCR, SMDRC, PCORR, or SCORR option in the TEST statement, PROC FREQ displays asymptotic tests for Gamma, Kendall’s Tau-b, Stuart’s Tau-c, Somers’ D(C|R), Somers’ D(R|C), the Pearson Correlation, or the Spearman Correlation, respectively. If you specify the MEASURES option in the TEST statement, PROC FREQ displays all these asymptotic tests. The test output includes the statistic, its asymptotic standard error (ASE), Confidence Limits, the ASE under the null hypothesis H0, the standardized test statistic (Z), and the one-sided and two-sided probability values.

• If you specify the KENTB, STUTC, SMDCR, SMDRC, PCORR, or SCORR option in the EXACT statement, PROC FREQ displays asymptotic and exact tests for the corresponding measure of association: Kendall’s Tau-b, Stuart’s Tau-c, Somers’ D(C|R), Somers’ D(R|C), the Pearson Correlation, or the Spearman correlation, respectively. The test output includes the correlation, its asymptotic standard error (ASE), Confidence Limits, the ASE under the null hypothesis H0, the standardized test statistic (Z), and the asymptotic and exact one-sided and two-sided probability values. If you also specify the POINT option in the EXACT statement, PROC FREQ displays the point probability for each exact test requested. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact Mid-p-Value for each test.

• If you specify the RISKDIFF option for 2 × 2 tables, PROC FREQ displays the Column 1 and Column 2 Risk Estimates. For each column, PROC FREQ displays the Row 1 Risk, Row 2 Risk, Total Risk, and Risk Difference, together with their asymptotic standard errors (ASE) and Asymptotic Confidence Limits. PROC FREQ also displays Exact Confidence Limits for the Row 1 Risk, Row 2 Risk, and Total Risk. If you specify the RISKDIFF option in the EXACT statement, PROC FREQ provides unconditional Exact Confidence Limits for the Risk Difference. You can suppress this table by specifying the RISKDIFF(NORISKS) option.

• If you specify the RISKDIFF(CL=) option for 2 × 2 tables, PROC FREQ displays the “Confidence Limits for the Proportion (Risk) Difference” table, which includes the Lower and Upper Confidence Limits for each confidence limit Type that you request (Exact, Hauck-Anderson, Miettinen-Nurminen score, Newcombe, or Wald).
If you request a noninferiority or superiority test for the proportion difference (RISKDIFF) by specifying the NONINF or SUP riskdiff-option, and if you specify METHOD=HA (Hauck-Anderson), METHOD=FM (Farrington-Manning score), or METHOD=WALD (Wald), PROC FREQ displays the following information: the Proportion Difference, the test ASE (H0, Sample, Sample H-A, or FM, depending on the method you specify), the test statistic Z, the probability value, the Noninferiority or Superiority Limit, and the test-based Confidence Limits. If you specify METHOD=NEWCOMBE (Newcombe), PROC FREQ displays the Proportion Difference, the Noninferiority or Superiority Limit, and the Newcombe Confidence Limits.

If you request an equivalence test for the proportion difference (RISKDIFF) by specifying the EQUIV riskdiff-option, and if you specify METHOD=HA (Hauck-Anderson), METHOD=FM (Farrington-Manning score), or METHOD=WALD (Wald), PROC FREQ displays the following information: the Proportion Difference and the test ASE (H0, Sample, Sample H-A, or FM, depending on the method you specify). PROC FREQ displays a two one-sided test (TOST) for equivalence, which includes test statistics (Z) and probability values for the Lower and Upper tests, together with the Overall probability value. PROC FREQ also displays the Equivalence Limits and the test-based Confidence Limits. If you specify METHOD=NEWCOMBE (Newcombe), PROC FREQ displays the Proportion Difference, the Equivalence Limits, and the Newcombe Confidence Limits.

If you request an equality test for the proportion difference (RISKDIFF) by specifying the EQUAL riskdiff-option, PROC FREQ displays the following information: the Proportion Difference and the test ASE (H0 or Sample), the test statistic Z, the One-Sided probability value (Pr > Z or Pr < Z), and the Two-Sided probability value, Pr > |Z|.

If you specify the MEASURES option or the RELRISK option for 2 × 2 tables, PROC FREQ displays the “Odds Ratio and Relative Risks” table, which includes the following statistics with their confidence limits: Odds Ratio, Relative Risk (Column 1), and Relative Risk (Column 2). If you specify the OR option in the EXACT statement, PROC FREQ also displays the “Exact Confidence Limits for the Odds Ratio” table. If you specify the RELRISK option in the EXACT statement, PROC FREQ displays the “Exact Confidence Limits for the Relative Risk” table.

If you specify the OR(CL=) option for 2 × 2 tables, PROC FREQ displays the “Confidence Limits for the Odds Ratio” table, which includes the Lower and Upper Confidence Limits for each confidence limit Type that you request (Exact, Score, or Wald).

If you specify the TREND option, PROC FREQ displays the Cochran-Armitage Trend Test for tables that are 2 × C or R × 2. For this test, PROC FREQ gives the Statistic (Z) and the one-sided and two-sided probability values. If you specify the TREND option in the EXACT statement, PROC FREQ also displays the exact one-sided and two-sided probability values for this test. If you specify the POINT option with the TREND option in the EXACT statement, PROC FREQ displays the exact point probability for the test statistic. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact Mid p-Value for the trend test.

If you specify the JT option, PROC FREQ displays the Jonckheere-Terpstra Test, showing the Statistic (JT), the standardized test statistic (Z), and the one-sided and two-sided probability values. If you specify the JT option in the EXACT statement, PROC FREQ also displays the exact one-sided and two-sided probability values for this test. If you specify the POINT option with the JT option in the EXACT statement, PROC FREQ displays the exact point probability for the test statistic. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact Mid p-Value for the Jonckheere-Terpstra test.
• If you specify the AGREE(PRINTKWTS) option, PROC FREQ displays the Kappa Coefficient Weights for square tables of dimension greater than 2.

• If you specify the AGREE option, for two-way tables PROC FREQ displays McNemar’s Test and the Simple Kappa Coefficient for $2 \times 2$ tables. For square tables larger than $2 \times 2$, PROC FREQ displays Bowker’s Test of Symmetry, the Simple Kappa Coefficient, and the Weighted Kappa Coefficient. For McNemar’s Test and Bowker’s Test of Symmetry, PROC FREQ displays the Statistic (S), the degrees of freedom (DF), and For the simple and weighted kappa coefficients, PROC FREQ displays the kappa values, asymptotic standard errors (ASE), and Confidence Limits. the probability value (Pr > S).

If you specify the MCNEM option in the EXACT statement, PROC FREQ also displays the exact probability value for McNemar’s test. If you specify the POINT option with the MCNEM option in the EXACT statement, PROC FREQ displays the exact point probability for the test statistic. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the Exact Mid $p$-Value for McNemar’s test.

• If you specify the KAPPA or WTKAP option in the TEST statement, PROC FREQ displays asymptotic tests for the simple kappa coefficient or the weighted kappa coefficient, respectively. If you specify the AGREE option in the TEST statement, PROC FREQ displays both these asymptotic tests. The test output includes the kappa coefficient, its asymptotic standard error (ASE), Confidence Limits, the ASE under the null hypothesis $H_0$, the standardized test statistic (Z), and the one-sided and two-sided probability values.

• If you specify the KAPPA or WTKAP option in the EXACT statement, PROC FREQ displays asymptotic and exact tests for the simple kappa coefficient or the weighted kappa coefficient, respectively. The test output includes the kappa coefficient, its asymptotic standard error (ASE), Confidence Limits, the ASE under the null hypothesis $H_0$, the standardized test statistic (Z), and the asymptotic and exact one-sided and two-sided probability values. If you specify the POINT option in the EXACT statement, PROC FREQ displays the point probability for each exact test requested. If you specify the MIDP option in the EXACT statement, PROC FREQ displays the exact Mid $p$-Value for each exact test.

• If you specify the MC option in the EXACT statement, PROC FREQ displays Monte Carlo estimates for all exact $p$-values requested by statistic-options in the EXACT statement. The Monte Carlo output includes the $p$-value Estimate, its Confidence Limits, the Number of Samples used to compute the Monte Carlo estimate, and the Initial Seed for random number generation.

• If you specify the AGREE option for a multiway square table, PROC FREQ displays Overall Simple and Weighted Kappa Coefficients, with their asymptotic standard errors (ASE) and Confidence Limits. PROC FREQ also displays Tests for Equal Kappa Coefficients, giving the Chi-Squares, degrees of freedom (DF), and probability values (Pr > ChiSq) for the Simple Kappa and Weighted Kappa. For multiple strata of $2 \times 2$ tables, PROC FREQ displays Cochran’s $Q$, giving the Statistic (Q), the degrees of freedom (DF), and the probability value (Pr > Q).

• If you specify the RISKDIFF(COMMON) option for a multiway $2 \times 2$ table, PROC FREQ displays the “Common Proportion (Risk) Difference” table, which includes the following information: Method (Mantel-Haenszel or Summary Score), Value of the common risk difference, Standard Error, and Confidence Limits.

• If you specify the CMH option, PROC FREQ displays Cochran-Mantel-Haenszel Statistics for the following three alternative hypotheses: Nonzero Correlation, Row Mean Scores Differ (ANOVA Statistic), and General Association. For each of these statistics, PROC FREQ gives the degrees of
freedom (DF) and the probability value (Prob). If you specify the MANTELFLEISS option, PROC FREQ displays the Mantel-Fleiss Criterion for $2 \times 2$ tables. For $2 \times 2$ tables, PROC FREQ also displays Estimates of the Common Relative Risk for Case-Control and Cohort studies, together with their confidence limits. These include both Mantel-Haenszel and Logit stratum-adjusted estimates of the common Odds Ratio, Column 1 Relative Risk, and Column 2 Relative Risk. Also for $2 \times 2$ tables, PROC FREQ displays the Breslow-Day Test for Homogeneity of the Odds Ratios. For this test, PROC FREQ gives the Chi-Square, the degrees of freedom (DF), and the probability value (Pr > ChiSq).

- If you specify the CMH option in the TABLES statement and also specify the COMOR option in the EXACT statement for a multiway $2 \times 2$ table, PROC FREQ displays exact confidence limits for the Common Odds Ratio. PROC FREQ also displays the Exact Test of H0: Common Odds Ratio = 1. The test output includes the Cell (1,1) Sum (S), Mean of S Under H0, One-sided Pr <= S, and Point Pr = S. PROC FREQ also provides exact two-sided probability values for the test, computed according to the following three methods: 2 * One-sided, Sum of probabilities <= Point probability, and Pr >= |S - Mean|. If you specify the MIDP option in the EXACT statement, PROC FREQ provides the exact Mid p-Value for the common odds ratio test.

- If you specify the CMH option in the TABLES statement and also specify the EQOR option in the EXACT statement for a multiway $2 \times 2$ table, PROC FREQ computes Zelen’s exact test for equal odds ratios. PROC FREQ displays Zelen’s test along with the asymptotic Breslow-Day test produced by the CMH option. PROC FREQ displays the test statistic, Zelen’s Exact Test (P), and the probability value, Exact Pr <= P.

- If you specify the GAILSIMON option in the TABLES statement for a multiway $2 \times 2$ tables, PROC FREQ displays the Gail-Simon test for qualitative interactions. The display include the following statistics and their p-values: Q+ (Positive Risk Differences), Q- (Negative Risk Differences), and Q (Two-Sided).

---

### ODS Table Names

PROC FREQ assigns a name to each table that it creates. You can use these names to refer to tables when you use the Output Delivery System (ODS) to select tables and create output data sets. For more information about ODS, see Chapter 20, “Using the Output Delivery System” (SAS/STAT User’s Guide).

Table 3.21 lists the ODS table names together with their descriptions and the options required to produce the tables. Note that the ALL option in the TABLES statement invokes the CHISQ, MEASURES, and CMH options.

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>BarnardsTest</td>
<td>Barnard’s exact test</td>
<td>EXACT</td>
<td>BARNARD</td>
</tr>
<tr>
<td>BinomialCLs</td>
<td>Binomial confidence limits</td>
<td>TABLES</td>
<td>BINOMIAL(AC</td>
</tr>
<tr>
<td>BinomialEquiv</td>
<td>Binomial equivalence analysis</td>
<td>TABLES</td>
<td>BINOMIAL(EQUIV)</td>
</tr>
<tr>
<td>BinomialEquivLimits</td>
<td>Binomial equivalence limits</td>
<td>TABLES</td>
<td>BINOMIAL(EQUIV)</td>
</tr>
<tr>
<td>BinomialEquivTest</td>
<td>Binomial equivalence test</td>
<td>TABLES</td>
<td>BINOMIAL(EQUIV)</td>
</tr>
<tr>
<td>BinomialNoninf</td>
<td>Binomial noninferiority test</td>
<td>TABLES</td>
<td>BINOMIAL(NONINF)</td>
</tr>
<tr>
<td>Binomial</td>
<td>Binomial proportion</td>
<td>TABLES</td>
<td>BINOMIAL</td>
</tr>
<tr>
<td>BinomialTest</td>
<td>Binomial proportion test</td>
<td>TABLES</td>
<td>BINOMIAL</td>
</tr>
<tr>
<td>ODS Table Name</td>
<td>Description</td>
<td>Statement</td>
<td>Option</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------------</td>
<td>-------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>BinomialSup</td>
<td>Binomial superiority test</td>
<td>TABLES</td>
<td>BINOMIAL(SUP)</td>
</tr>
<tr>
<td>BreslowDayTest</td>
<td>Breslow-Day test</td>
<td>TABLES</td>
<td>CMH</td>
</tr>
<tr>
<td>CMH</td>
<td>Cochran-Mantel-Haenszel statistics</td>
<td>TABLES</td>
<td>CMH</td>
</tr>
<tr>
<td>ChiSq</td>
<td>Chi-square tests</td>
<td>TABLES</td>
<td>CHISQ</td>
</tr>
<tr>
<td>CochransQ</td>
<td>Cochran’s Q</td>
<td>TABLES</td>
<td>AGREE</td>
</tr>
<tr>
<td>ColScores</td>
<td>Column scores</td>
<td>TABLES</td>
<td>SCOROUT</td>
</tr>
<tr>
<td>CommonOdds-RatioCl</td>
<td>Exact confidence limits for the common odds ratio</td>
<td>EXACT</td>
<td>COMOR</td>
</tr>
<tr>
<td>CommonOdds-RatioTest</td>
<td>Common odds ratio exact test</td>
<td>EXACT</td>
<td>COMOR</td>
</tr>
<tr>
<td>CommonPdiff</td>
<td>Common proportion difference</td>
<td>TABLES</td>
<td>RISKDIFF(COMMON)</td>
</tr>
<tr>
<td>CommonRelRisks</td>
<td>Common relative risks</td>
<td>TABLES</td>
<td>CMH</td>
</tr>
<tr>
<td>CrossList</td>
<td>Crosstabulation table in column format</td>
<td>TABLES</td>
<td>CROSSLIST</td>
</tr>
<tr>
<td>CrossTabFreqs</td>
<td>Crosstabulation table</td>
<td>TABLES</td>
<td>(n-way table, n &gt; 1)</td>
</tr>
<tr>
<td>EqualKappaTest</td>
<td>Test for equal simple kappas</td>
<td>TABLES</td>
<td>AGREE</td>
</tr>
<tr>
<td>EqualKappaTests</td>
<td>Tests for equal kappas</td>
<td>TABLES</td>
<td>AGREE</td>
</tr>
<tr>
<td>EqualOddsRatios</td>
<td>Tests for equal odds ratios</td>
<td>EXACT</td>
<td>EQOR</td>
</tr>
<tr>
<td>GailSimon</td>
<td>Gail-Simon test</td>
<td>TABLES</td>
<td>GAILSIMON</td>
</tr>
<tr>
<td>FishersExact</td>
<td>Fisher’s exact test</td>
<td>EXACT</td>
<td>FISHER</td>
</tr>
<tr>
<td>FishersExactMC</td>
<td>Monte Carlo estimates for Fisher’s exact test</td>
<td>EXACT</td>
<td>FISHER / MC</td>
</tr>
<tr>
<td>Gamma</td>
<td>Gamma</td>
<td>TEST</td>
<td>GAMMA</td>
</tr>
<tr>
<td>GammaTest</td>
<td>Gamma test</td>
<td>TEST</td>
<td>GAMMA</td>
</tr>
<tr>
<td>JTTest</td>
<td>Jonckheere-Terpstra test</td>
<td>TABLES</td>
<td>JT</td>
</tr>
<tr>
<td>JTTestMC</td>
<td>Monte Carlo estimates for Jonckheere-Terpstra exact test</td>
<td>EXACT</td>
<td>JT / MC</td>
</tr>
<tr>
<td>Kappa</td>
<td>Simple kappa coefficient</td>
<td>TEST</td>
<td>KAPPA</td>
</tr>
<tr>
<td>KappaMC</td>
<td>Monte Carlo exact test for simple kappa coefficient</td>
<td>EXACT</td>
<td>KAPPA / MC</td>
</tr>
<tr>
<td>KappaStatistics</td>
<td>Kappa statistics</td>
<td>TABLES</td>
<td>AGREE, no TEST or EXACT (r x r table, r &gt; 2)</td>
</tr>
<tr>
<td>ODS Table Name</td>
<td>Description</td>
<td>Statement</td>
<td>Option</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>KappaTest</td>
<td>Simple kappa test</td>
<td>TEST or EXACT</td>
<td>KAPPA</td>
</tr>
<tr>
<td>KappaWeights</td>
<td>Kappa weights</td>
<td>TABLES</td>
<td>AGREE(PRINTKWTS)</td>
</tr>
<tr>
<td>List</td>
<td>List format multiway table</td>
<td>TABLES</td>
<td>LIST</td>
</tr>
<tr>
<td>LRChiSq</td>
<td>Likelihood ratio chi-square exact test</td>
<td>EXACT</td>
<td>LRCHI</td>
</tr>
<tr>
<td>LRChiSqMC</td>
<td>Monte Carlo exact test for likelihood ratio chi-square</td>
<td>EXACT</td>
<td>LRCHI / MC</td>
</tr>
<tr>
<td>MantelFleiss</td>
<td>Mantel-Fleiss criterion</td>
<td>TABLES</td>
<td>CMH(MF) (h x 2 x 2 table)</td>
</tr>
<tr>
<td>McNemarsTest</td>
<td>McNemar’s test</td>
<td>TABLES</td>
<td>AGREE (2 x 2 table)</td>
</tr>
<tr>
<td>Measures</td>
<td>Measures of association</td>
<td>TABLES</td>
<td>MEASURES</td>
</tr>
<tr>
<td>MHChiSq</td>
<td>Mantel-Haenszel chi-square exact test</td>
<td>EXACT</td>
<td>MHCHI</td>
</tr>
<tr>
<td>MHChiSqMC</td>
<td>Monte Carlo exact test for Mantel-Haenszel chi-square</td>
<td>EXACT</td>
<td>MHCHI / MC</td>
</tr>
<tr>
<td>NLevels</td>
<td>Number of variable levels</td>
<td>PROC or TABLES</td>
<td>NLEVELS</td>
</tr>
<tr>
<td>OddsRatioCLs</td>
<td>Odds ratio confidence limits</td>
<td>TABLES</td>
<td>OR(CL=) (2 x 2 table)</td>
</tr>
<tr>
<td>OddsRatioExactCL</td>
<td>Exact confidence limits</td>
<td>EXACT</td>
<td>OR (2 x 2 table)</td>
</tr>
<tr>
<td>OneWayChiSq</td>
<td>One-way chi-square test</td>
<td>TABLES</td>
<td>CHISQ (one-way table)</td>
</tr>
<tr>
<td>OneWayChiSqMC</td>
<td>Monte Carlo exact test for one-way chi-square</td>
<td>EXACT</td>
<td>CHISQ / MC (one-way table)</td>
</tr>
<tr>
<td>OneWayFreqs</td>
<td>One-way frequencies</td>
<td>PROC or TABLES</td>
<td>(no TABLES stmt) (one-way table)</td>
</tr>
<tr>
<td>OneWayLRChiSq</td>
<td>One-way likelihood ratio chi-square test</td>
<td>TABLES</td>
<td>CHISQ(LRCHI) (one-way table)</td>
</tr>
<tr>
<td>OverallKappa</td>
<td>Overall simple kappa</td>
<td>TABLES</td>
<td>AGREE (h x 2 x 2 table)</td>
</tr>
<tr>
<td>OverallKappas</td>
<td>Overall kappa coefficients</td>
<td>TABLES</td>
<td>AGREE (h x r x r, r &gt; 2)</td>
</tr>
<tr>
<td>PdiffCLs</td>
<td>Proportion difference confidence limits</td>
<td>TABLES</td>
<td>RISKDIFF(CL=) (2 x 2 table)</td>
</tr>
<tr>
<td>PdiffEquiv</td>
<td>Equivalence analysis for the proportion difference</td>
<td>TABLES</td>
<td>RISKDIFF(EQUIV) (2 x 2 table)</td>
</tr>
<tr>
<td>PdiffEquivLimits</td>
<td>Equivalence limits for the proportion difference</td>
<td>TABLES</td>
<td>RISKDIFF(EQUIV) (2 x 2 table)</td>
</tr>
<tr>
<td>PdiffEquivTest</td>
<td>Equivalence test for the proportion difference</td>
<td>TABLES</td>
<td>RISKDIFF(EQUIV) (2 x 2 table)</td>
</tr>
<tr>
<td>PdiffNoninf</td>
<td>Noninferiority test for the proportion difference</td>
<td>TABLES</td>
<td>RISKDIFF(NONINF) (2 x 2 table)</td>
</tr>
<tr>
<td>PdiffSup</td>
<td>Superiority test for the proportion difference</td>
<td>TABLES</td>
<td>RISKDIFF(SUP) (2 x 2 table)</td>
</tr>
</tbody>
</table>
### Table 3.21 continued

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>PdiffTest</td>
<td>Proportion difference test</td>
<td>TABLES</td>
<td>RISKDIFF(EQUAL) (2 x 2 table)</td>
</tr>
<tr>
<td>PearsonChiSq</td>
<td>Pearson chi-square exact test</td>
<td>EXACT</td>
<td>PCHI</td>
</tr>
<tr>
<td>PearsonChiSqMC</td>
<td>Monte Carlo exact test for Pearson chi-square</td>
<td>EXACT</td>
<td>PCHI / MC</td>
</tr>
<tr>
<td>PearsonCorr</td>
<td>Pearson correlation</td>
<td>TEST</td>
<td>PCORR</td>
</tr>
<tr>
<td>PearsonCorrMC</td>
<td>Monte Carlo exact test for Pearson correlation</td>
<td>EXACT</td>
<td>PCORR / MC</td>
</tr>
<tr>
<td>PearsonCorrTest</td>
<td>Pearson correlation test</td>
<td>TEST</td>
<td>PCORR</td>
</tr>
<tr>
<td>PlCorr</td>
<td>Polychoric correlation</td>
<td>TEST</td>
<td>PLCORR</td>
</tr>
<tr>
<td>PlCorrTest</td>
<td>Polychoric correlation test</td>
<td>TEST</td>
<td>PLCORR</td>
</tr>
<tr>
<td>RelativeRisks</td>
<td>Relative risk estimates</td>
<td>TABLES</td>
<td>RELRISK (2 x 2 table)</td>
</tr>
<tr>
<td>RelRisk1ExactCL</td>
<td>Exact confidence limits for column 1 relative risk</td>
<td>EXACT</td>
<td>RELRISK (2 x 2 table)</td>
</tr>
<tr>
<td>RelRisk2ExactCL</td>
<td>Exact confidence limits for column 2 relative risk</td>
<td>EXACT</td>
<td>RELRISK (2 x 2 table)</td>
</tr>
<tr>
<td>RiskDiffCol1</td>
<td>Column 1 risk estimates</td>
<td>TABLES</td>
<td>RISKDIFF (2 x 2 table)</td>
</tr>
<tr>
<td>RiskDiffCol2</td>
<td>Column 2 risk estimates</td>
<td>TABLES</td>
<td>RISKDIFF (2 x 2 table)</td>
</tr>
<tr>
<td>RowScores</td>
<td>Row scores</td>
<td>TABLES</td>
<td>SCOROUT</td>
</tr>
<tr>
<td>SomersDCR</td>
<td>Somers’ $D(C \mid R)$</td>
<td>TEST</td>
<td>SMDCR</td>
</tr>
<tr>
<td>SomersDCRTest</td>
<td>Somers’ $D(C \mid R)$ test</td>
<td>TEST</td>
<td>SMDCR</td>
</tr>
<tr>
<td>SomersDRC</td>
<td>Somers’ $D(R \mid C)$</td>
<td>TEST</td>
<td>SMDRC</td>
</tr>
<tr>
<td>SomersDRCMC</td>
<td>Monte Carlo exact test for Somers’ $D(R \mid C)$</td>
<td>EXACT</td>
<td>SMDRC / MC</td>
</tr>
<tr>
<td>SomersDRCTest</td>
<td>Somers’ $D(R \mid C)$ test</td>
<td>TEST</td>
<td>SMDRC</td>
</tr>
<tr>
<td>SpearmanCorr</td>
<td>Spearman correlation</td>
<td>TEST</td>
<td>SCORR</td>
</tr>
<tr>
<td>SpearmanCorrMC</td>
<td>Monte Carlo exact test for Spearman correlation</td>
<td>EXACT</td>
<td>SCORR / MC</td>
</tr>
<tr>
<td>SpearmanCorrTest</td>
<td>Spearman correlation test</td>
<td>TEST</td>
<td>SCORR</td>
</tr>
<tr>
<td>SymmetryTest</td>
<td>Test of symmetry</td>
<td>TABLES</td>
<td>AGREE</td>
</tr>
</tbody>
</table>
Table 3.21 continued

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>TauB</td>
<td>Kendall’s tau-(b)</td>
<td>TEST</td>
<td>KENTB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or EXACT</td>
<td>KENTB</td>
</tr>
<tr>
<td>TauBMC</td>
<td>Monte Carlo exact test for Kendall’s tau-(b)</td>
<td>EXACT</td>
<td>KENTB / MC</td>
</tr>
<tr>
<td>TauBTest</td>
<td>Kendall’s tau-(b) test</td>
<td>TEST</td>
<td>KENTB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or EXACT</td>
<td>KENTB</td>
</tr>
<tr>
<td>TauC</td>
<td>Stuart’s tau-(c)</td>
<td>TEST</td>
<td>STUTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or EXACT</td>
<td>STUTC</td>
</tr>
<tr>
<td>TauCMC</td>
<td>Monte Carlo exact test for Stuart’s tau-(c)</td>
<td>EXACT</td>
<td>STUTC / MC</td>
</tr>
<tr>
<td>TauCTest</td>
<td>Stuart’s tau-(c) test</td>
<td>TEST</td>
<td>STUTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or EXACT</td>
<td>STUTC</td>
</tr>
<tr>
<td>TrendTest</td>
<td>Cochran-Armitage trend test</td>
<td>TABLES</td>
<td>TREND</td>
</tr>
<tr>
<td>TrendTestMC</td>
<td>Monte Carlo exact test for trend</td>
<td>EXACT</td>
<td>TREND / MC</td>
</tr>
<tr>
<td>WtKappa</td>
<td>Weighted kappa coefficient</td>
<td>TEST</td>
<td>WTKAP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or EXACT</td>
<td>WTKAP</td>
</tr>
<tr>
<td>WtKappaMC</td>
<td>Monte Carlo exact test for weighted kappa coefficient</td>
<td>EXACT</td>
<td>WTKAP / MC</td>
</tr>
<tr>
<td>WtKappaTest</td>
<td>Weighted kappa test</td>
<td>TEST</td>
<td>WTKAP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or EXACT</td>
<td>WTKAP</td>
</tr>
</tbody>
</table>

**ODS Graphics**


Before you create graphs, ODS Graphics must be enabled (for example, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” in that chapter.

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “A Primer on ODS Statistical Graphics” in that chapter.

When ODS Graphics is enabled, you can request specific plots with the PLOTS= option in the TABLES statement. To produce a frequency plot or cumulative frequency plot, you must specify the FREQPLOT or CUMFREQPLOT plot-request, respectively, in the PLOTS= option. To produce a mosaic plot, you must specify the MOSAICPLOT plot-request in the PLOTS= option. You can also produce frequency, cumulative frequency, and mosaic plots by specifying the PLOTS=ALL option. By default, PROC FREQ produces all other plots that are associated with the analyses that you request in the TABLES statement. You can suppress the default plots and request specific plots by using the PLOTS(ONLY)= option. See the description of the PLOTS= option for details.

PROC FREQ assigns a name to each graph that it creates with ODS Graphics. You can use these names to refer to the graphs. Table 3.22 lists the names of the graphs that PROC FREQ generates together with their
Table 3.22 Graphs Produced by PROC FREQ

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Description</th>
<th>PLOTS= Option</th>
<th>TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgreePlot</td>
<td>Agreement plot</td>
<td>AGREEPLOT</td>
<td>AGREE (r × r table)</td>
</tr>
<tr>
<td>CumFreqPlot</td>
<td>Cumulative frequency plot</td>
<td>CUMFREQPLOT</td>
<td>One-way table request</td>
</tr>
<tr>
<td>DeviationPlot</td>
<td>Deviation plot</td>
<td>DEVIATIONPLOT</td>
<td>CHISQ (one-way table)</td>
</tr>
<tr>
<td>FreqPlot</td>
<td>Frequency plot</td>
<td>FREQPLOT</td>
<td>Any table request</td>
</tr>
<tr>
<td>KappaPlot</td>
<td>Kappa plot</td>
<td>KAPPAPLOT</td>
<td>AGREE (h × r × r table)</td>
</tr>
<tr>
<td>MosaicPlot</td>
<td>Mosaic plot</td>
<td>MOSAICPLOT</td>
<td>Two-way or multiway table request</td>
</tr>
<tr>
<td>ORPlot</td>
<td>Odds ratio plot</td>
<td>ODDSRAIOTPLOT</td>
<td>MEASURES or RELRISK (h × 2 × 2 table)</td>
</tr>
<tr>
<td>RelRiskPlot</td>
<td>Relative risk plot</td>
<td>RELRISKPLOT</td>
<td>MEASURES or RELRISK (h × 2 × 2 table)</td>
</tr>
<tr>
<td>RiskDiffPlot</td>
<td>Risk difference plot</td>
<td>RISKDIFFPLOT</td>
<td>RISKDIFF (h × 2 × 2 table)</td>
</tr>
<tr>
<td>WtKappaPlot</td>
<td>Weighted kappa plot</td>
<td>WTKAPPAPLOT</td>
<td>AGREE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(h × r × r table, r &gt; 2)</td>
</tr>
</tbody>
</table>

Examples: FREQ Procedure

Example 3.1: Output Data Set of Frequencies

The eye and hair color of children from two different regions of Europe are recorded in the data set Color. Instead of recording one observation per child, the data are recorded as cell counts, where the variable Count contains the number of children exhibiting each of the 15 eye and hair color combinations. The data set does not include missing combinations.

The following DATA step statements create the SAS data set Color:

```sas
data Color;
  input Region Eyes $ Hair $ Count @@;
  label Eyes = 'Eye Color'
  Hair = 'Hair Color'
  Region = 'Geographic Region';
datalines;
1 blue fair 23 1 blue red 7 1 blue medium 24
1 blue dark 11 1 green fair 19 1 green red 7
1 green medium 18 1 green dark 14 1 brown fair 34
1 brown red 5 1 brown medium 41 1 brown dark 40
1 brown black 3 2 blue fair 46 2 blue red 21
2 blue medium 44 2 blue dark 40 2 blue black 6
2 green fair 50 2 green red 31 2 green medium 37
2 green dark 23 2 brown fair 56 2 brown red 42
2 brown medium 53 2 brown dark 54 2 brown black 13
;
```

The following PROC FREQ statements read the Color data set and create an output data set that contains the frequencies, percentages, and expected cell frequencies of the two-way table of Eyes by Hair. The TABLES
Chapter 3: The FREQ Procedure

The statement requests three tables: a frequency table for Eyes, a frequency table for Hair, and a crosstabulation table for Eyes by Hair. The OUT= option creates the FreqCount data set, which contains the crosstabulation table frequencies. The OUTEXPECT option outputs the expected table cell frequencies to FreqCount, and the SPARSE option includes zero cell frequencies in the output data set. The WEIGHT statement specifies that the variable Count contains the observation weights. These statements create Output 3.1.1 through Output 3.1.3.

```plaintext
proc freq data=Color;
    tables Eyes Hair Eyes*Hair / out=FreqCount outexpect sparse;
    weight Count;
    title 'Eye and Hair Color of European Children';
run;

proc print data=FreqCount noobs;
    title2 'Output Data Set from PROC FREQ';
run;
```

Output 3.1.1 displays the two frequency tables produced by PROC FREQ: one showing the distribution of eye color, and one showing the distribution of hair color. By default, PROC FREQ lists the variables values in alphabetical order. The ‘Eyes*Hair’ specification produces a crosstabulation table, shown in Output 3.1.2, with eye color defining the table rows and hair color defining the table columns. A zero cell frequency for green eyes and black hair indicates that this eye and hair color combination does not occur in the data.

The output data set FreqCount (Output 3.1.3) contains frequency counts and percentages for the last table requested in the TABLES statement, Eyes by Hair. Because the SPARSE option is specified, the data set includes the observation with a zero frequency. The variable Expected contains the expected frequencies, as requested by the OUTEXPECT option.

### Output 3.1.1 Frequency Tables

<table>
<thead>
<tr>
<th>Eye Color</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>222</td>
<td>29.13</td>
<td>222</td>
<td>29.13</td>
</tr>
<tr>
<td>brown</td>
<td>341</td>
<td>44.75</td>
<td>563</td>
<td>73.88</td>
</tr>
<tr>
<td>green</td>
<td>199</td>
<td>26.12</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hair Color</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>22</td>
<td>2.89</td>
<td>22</td>
<td>2.89</td>
</tr>
<tr>
<td>dark</td>
<td>182</td>
<td>23.88</td>
<td>204</td>
<td>26.77</td>
</tr>
<tr>
<td>fair</td>
<td>228</td>
<td>29.92</td>
<td>432</td>
<td>56.69</td>
</tr>
<tr>
<td>medium</td>
<td>217</td>
<td>28.48</td>
<td>649</td>
<td>85.17</td>
</tr>
<tr>
<td>red</td>
<td>113</td>
<td>14.83</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>
**Example 3.1: Output Data Set of Frequencies**

### Output 3.1.2  Crosstabulation Table

<table>
<thead>
<tr>
<th>Eyes (Eye Color)</th>
<th>Hair (Hair Color)</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>black</td>
<td>6</td>
<td>0.79</td>
<td>2.70</td>
<td>27.27</td>
</tr>
<tr>
<td></td>
<td>dark</td>
<td>51</td>
<td>0.79</td>
<td>22.97</td>
<td>22.97</td>
</tr>
<tr>
<td></td>
<td>fair</td>
<td>69</td>
<td>0.89</td>
<td>31.08</td>
<td>31.08</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>68</td>
<td>0.89</td>
<td>30.63</td>
<td>30.63</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>28</td>
<td>0.36</td>
<td>12.61</td>
<td>12.61</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>222</td>
<td>29.13</td>
<td>24.78</td>
<td>24.78</td>
</tr>
</tbody>
</table>

| brown           | black             | 16        | 2.10    | 4.69   | 72.73  |
|                 | dark              | 94        | 12.34   | 27.57  | 44.75  |
|                 | fair              | 90        | 12.34   | 27.57  | 44.75  |
|                 | medium            | 94        | 12.34   | 27.57  | 44.75  |
|                 | red               | 47        | 6.17    | 13.78  | 13.78  |
|                 | Total             | 341       | 44.75   | 41.59  | 41.59  |

| green           | black             | 0         | 0.00    | 0.00   | 0.00   |
|                 | dark              | 37        | 4.86    | 18.59  | 30.26  |
|                 | fair              | 69        | 9.06    | 34.67  | 51.65  |
|                 | medium            | 55        | 7.22    | 27.64  | 43.32  |
|                 | red               | 38        | 4.99    | 19.10  | 33.63  |
|                 | Total             | 199       | 26.12   | 33.63  | 33.63  |

### Total

<table>
<thead>
<tr>
<th>Eyes (Eye Color)</th>
<th>Hair (Hair Color)</th>
<th>COUNT</th>
<th>EXPECTED</th>
<th>PERCENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>black</td>
<td>6</td>
<td>6.409</td>
<td>0.7874</td>
</tr>
<tr>
<td>blue</td>
<td>dark</td>
<td>51</td>
<td>53.024</td>
<td>6.6929</td>
</tr>
<tr>
<td>blue</td>
<td>fair</td>
<td>69</td>
<td>66.425</td>
<td>9.0551</td>
</tr>
<tr>
<td>blue</td>
<td>medium</td>
<td>68</td>
<td>63.220</td>
<td>8.9239</td>
</tr>
<tr>
<td>blue</td>
<td>red</td>
<td>28</td>
<td>32.921</td>
<td>3.6745</td>
</tr>
<tr>
<td>brown</td>
<td>black</td>
<td>16</td>
<td>9.845</td>
<td>2.0997</td>
</tr>
<tr>
<td>brown</td>
<td>dark</td>
<td>94</td>
<td>81.446</td>
<td>12.3360</td>
</tr>
<tr>
<td>brown</td>
<td>fair</td>
<td>90</td>
<td>102.031</td>
<td>11.8110</td>
</tr>
<tr>
<td>brown</td>
<td>medium</td>
<td>94</td>
<td>97.109</td>
<td>12.3360</td>
</tr>
<tr>
<td>brown</td>
<td>red</td>
<td>47</td>
<td>50.568</td>
<td>6.1680</td>
</tr>
<tr>
<td>green</td>
<td>black</td>
<td>0</td>
<td>5.745</td>
<td>0.0000</td>
</tr>
<tr>
<td>green</td>
<td>dark</td>
<td>37</td>
<td>47.530</td>
<td>4.8556</td>
</tr>
<tr>
<td>green</td>
<td>fair</td>
<td>69</td>
<td>59.543</td>
<td>9.0551</td>
</tr>
<tr>
<td>green</td>
<td>medium</td>
<td>55</td>
<td>56.671</td>
<td>7.2178</td>
</tr>
<tr>
<td>green</td>
<td>red</td>
<td>38</td>
<td>29.510</td>
<td>4.9869</td>
</tr>
</tbody>
</table>
Example 3.2: Frequency Dot Plots

This example produces frequency dot plots for the children’s eye and hair color data from Example 3.1.

PROC FREQ produces plots by using ODS Graphics to create graphs as part of the procedure output. Frequency plots are available for any frequency or crosstabulation table request. You can display frequency plots as bar charts or dot plots. You can use plot-options to specify the orientation (vertical or horizontal), scale, and layout of the plots.

The following PROC FREQ statements request frequency tables and dot plots. The first TABLES statement requests a one-way frequency table of Hair and a crosstabulation table of Eyes by Hair. The PLOTS= option requests frequency plots for the tables, and the TYPE=DOTPLOT plot-option specifies dot plots. By default, frequency plots are produced as bar charts. ODS Graphics must be enabled before producing plots.

The second TABLES statement requests a crosstabulation table of Region by Hair and a frequency dot plot for this table. The SCALE=PERCENT plot-option plots percentages instead of frequency counts. SCALE=LOG and SCALE=SQRT plot-options are also available to plot log frequencies and square roots of frequencies, respectively.

The ORDER=FREQ option in the PROC FREQ statement orders the variable levels by frequency. This order applies to the frequency and crosstabulation table displays and also to the corresponding frequency plots.

ods graphics on;
proc freq data=Color order=freq;
  tables Hair Hair*Eyes / plots=freqplot(type=dotplot);
  tables Hair*Region / plots=freqplot(type=dotplot scale=percent);
  weight Count;
  title 'Eye and Hair Color of European Children';
run;
ods graphics off;

Output 3.2.1, Output 3.2.2, and Output 3.2.3 display the dot plots produced by PROC FREQ. By default, the orientation of dot plots is horizontal, which places the variable levels on the Y axis. You can specify the ORIENT=VERTICAL plot-option to request a vertical orientation. For two-way plots, you can use the TWOWAY= plot-option to specify the plot layout. The default layout (shown in Output 3.2.2 and Output 3.2.3) is GROUPVERTICAL. Two-way layouts STACKED and GROUPHORIZONTAL are also available.
Output 3.2.1 One-Way Frequency Dot Plot
Output 3.2.2 Two-Way Frequency Dot Plot

Distribution of Hair by Eyes

<table>
<thead>
<tr>
<th>Eye Color = brown</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
</tr>
<tr>
<td>red</td>
</tr>
<tr>
<td>dark</td>
</tr>
<tr>
<td>medium</td>
</tr>
<tr>
<td>fair</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eye Color = blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
</tr>
<tr>
<td>red</td>
</tr>
<tr>
<td>dark</td>
</tr>
<tr>
<td>medium</td>
</tr>
<tr>
<td>fair</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eye Color = green</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
</tr>
<tr>
<td>red</td>
</tr>
<tr>
<td>dark</td>
</tr>
<tr>
<td>medium</td>
</tr>
<tr>
<td>fair</td>
</tr>
</tbody>
</table>

Frequency
Example 3.3: Chi-Square Goodness-of-Fit Tests

This example examines whether the children’s hair color (from Example 3.1) has a specified multinomial distribution for the two geographical regions. The hypothesized distribution of hair color is 30% fair, 12% red, 30% medium, 25% dark, and 3% black.

In order to test the hypothesis for each region, the data are first sorted by Region. Then the FREQ procedure uses a BY statement to produce a separate table for each BY group (Region). The option ORDER=DATA orders the variable values (hair color) in the frequency table by their order in the input data set. The TABLES statement requests a frequency table for hair color, and the option NOCUM suppresses the display of the cumulative frequencies and percentages.

The CHISQ option requests a chi-square goodness-of-fit test for the frequency table of Hair. The TESTP= option specifies the hypothesized (or test) percentages for the chi-square test; the number of percentages listed equals the number of table levels, and the percentages sum to 100%. The TESTP= percentages are listed in the same order as the corresponding variable levels appear in frequency table.

The PLOTS= option requests a deviation plot, which is associated with the CHISQ option and displays the relative deviations from the test frequencies. The TYPE=DOTPLOT plot-option requests a dot plot instead.

Output 3.2.3 Two-Way Percent Dot Plot

Distribution of Hair by Region

Geographic Region = 2

Geographic Region = 1

Hair Color

black
red
dark
medium
fair

Percent

0 5 10 15 20
of the default type, which is a bar chart. ODS Graphics must be enabled before producing plots. These statements produce Output 3.3.1 through Output 3.3.4.

```proc sort data=Color;
   by Region;
run;
ods graphics on;
proc freq data=Color order=data;
   tables Hair / nocum chisq testp=(30 12 30 25 3)
      plots(only)=deviationplot(type=dotplot);
   weight Count;
   by Region;
   title 'Hair Color of European Children';
run;
ods graphics off;
```

**Output 3.3.1** Frequency Table and Chi-Square Test for Region 1

<table>
<thead>
<tr>
<th>Hair Color of European Children</th>
<th>Geographic Region=1</th>
<th>-----------------------------</th>
<th>The FREQ Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hair Color</td>
<td></td>
<td></td>
<td>Test Percent</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------</td>
<td>-----------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>fair</td>
<td>76</td>
<td>30.89</td>
<td>30.00</td>
</tr>
<tr>
<td>red</td>
<td>19</td>
<td>7.72</td>
<td>12.00</td>
</tr>
<tr>
<td>medium</td>
<td>83</td>
<td>33.74</td>
<td>30.00</td>
</tr>
<tr>
<td>dark</td>
<td>65</td>
<td>26.42</td>
<td>25.00</td>
</tr>
<tr>
<td>black</td>
<td>3</td>
<td>1.22</td>
<td>3.00</td>
</tr>
</tbody>
</table>

**Chi-Square Test** for Specified Proportions

<table>
<thead>
<tr>
<th>Chi-Square Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>7.7602</td>
</tr>
<tr>
<td>DF</td>
<td>4</td>
</tr>
<tr>
<td>Pr &gt; ChiSq</td>
<td>0.1008</td>
</tr>
</tbody>
</table>

Output 3.3.1 shows the frequency table and chi-square test for Region 1. The frequency table lists the variable values (hair color) in the order in which they appear in the data set. The “Test Percent” column lists the hypothesized percentages for the chi-square test. Always check that you have ordered the TESTP= percentages to correctly match the order of the variable levels.
Output 3.3.2 shows the deviation plot for Region 1, which displays the relative deviations from the hypothesized values. The relative deviation for a level is the difference between the observed and hypothesized (test) percentage divided by the test percentage. You can suppress the chi-square $p$-value that is displayed by default in the deviation plot by specifying the NOSTATS plot-option.

**Output 3.3.2** Deviation Plot for Region 1

Output 3.3.3 and Output 3.3.4 show the results for Region 2. PROC FREQ computes a chi-square statistic for each region. The chi-square statistic is significant at the 0.05 level for Region 2 ($p=0.0003$) but not for Region 1. This indicates a significant departure from the hypothesized percentages in Region 2.
### Output 3.3.3  Frequency Table and Chi-Square Test for Region 2

#### Hair Color of European Children

---

**Geographic Region=2**

The FREQ Procedure

<table>
<thead>
<tr>
<th>Hair</th>
<th>Frequency</th>
<th>Percent</th>
<th>Test Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>fair</td>
<td>152</td>
<td>29.46</td>
<td>30.00</td>
</tr>
<tr>
<td>red</td>
<td>94</td>
<td>18.22</td>
<td>12.00</td>
</tr>
<tr>
<td>medium</td>
<td>134</td>
<td>25.97</td>
<td>30.00</td>
</tr>
<tr>
<td>dark</td>
<td>117</td>
<td>22.67</td>
<td>25.00</td>
</tr>
<tr>
<td>black</td>
<td>19</td>
<td>3.68</td>
<td>3.00</td>
</tr>
</tbody>
</table>

---

**Chi-Square Test**

for Specified Proportions

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>21.3824</td>
</tr>
<tr>
<td>DF</td>
<td>4</td>
</tr>
<tr>
<td>Pr &gt; ChiSq</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
Example 3.4: Binomial Proportions

In this example, PROC FREQ computes binomial proportions, confidence limits, and tests. The example uses the eye and hair color data from Example 3.1. By default, PROC FREQ computes the binomial proportion as the proportion of observations in the first level of the one-way table. You can designate a different level by using the LEVEL=binomial-option.

The following PROC FREQ statements compute the proportion of children with brown eyes (from the data set in Example 3.1) and test the null hypothesis that the population proportion equals 50%. These statements also compute an equivalence for the proportion of children with fair hair.

The first TABLES statement requests a one-way frequency table for the variable Eyes. The BINOMIAL option requests the binomial proportion, confidence limits, and test. PROC FREQ computes the proportion with Eyes = ‘brown’, which is the first level displayed in the table. The AC, WILSON, and EXACT binomial-options request the following confidence limits types: Agresti-Coull, Wilson (score), and exact (Clopper-Pearson). By default, PROC FREQ provides Wald and exact (Clopper-Pearson) confidence limits for the binomial proportion. The BINOMIAL option also produces an asymptotic Wald test that the proportion equals 0.5. You can specify a different test proportion with the P=binomial-option. The ALPHA=0.1 option specifies that $\alpha = 10\%$, which produces 90% confidence limits.
Chapter 3: The FREQ Procedure

The second TABLES statement requests a one-way frequency table for the variable Hair. The BINOMIAL option requests the proportion for the first level, Hair = 'fair'. The EQUIV binomial-option requests an equivalence test for the binomial proportion. The P=.28 option specifies 0.28 as the null hypothesis proportion, and the MARGIN=.1 option specifies 0.1 as the equivalence test margin.

```sas
proc freq data=Color order=freq;
  tables Eyes / binomial(ac wilson exact) alpha=.1;
  tables Hair / binomial(equiv p=.28 margin=.1);
  weight Count;
  title 'Hair and Eye Color of European Children';
run;
```

Output 3.4.1 displays the results for eye color, and Output 3.4.2 displays the results for hair color.

**Output 3.4.1** Binomial Proportion for Eye Color

```
Hair and Eye Color of European Children

The FREQ Procedure

Eye Color

<table>
<thead>
<tr>
<th>Eyes</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown</td>
<td>341</td>
<td>44.75</td>
<td>341</td>
<td>44.75</td>
</tr>
<tr>
<td>blue</td>
<td>222</td>
<td>29.13</td>
<td>563</td>
<td>73.88</td>
</tr>
<tr>
<td>green</td>
<td>199</td>
<td>26.12</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Binomial Proportion
Eyes = brown

Proportion 0.4475
ASE 0.0180

Type 90% Confidence Limits

<table>
<thead>
<tr>
<th>Type</th>
<th>Wilson</th>
<th>Agresti-Coull</th>
<th>Clopper-Pearson (Exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilson</td>
<td>0.4181</td>
<td>0.4181</td>
<td>0.4174</td>
</tr>
<tr>
<td>Agresti-Coull</td>
<td>0.4181</td>
<td>0.4181</td>
<td>0.4174</td>
</tr>
<tr>
<td>Clopper-Pearson (Exact)</td>
<td>0.4174</td>
<td>0.4174</td>
<td>0.4174</td>
</tr>
</tbody>
</table>

Test of H0: Proportion = 0.5

| ASE under H0 | Z   | One-sided Pr < Z | Two-sided Pr > |Z| |
|--------------|-----|------------------|----------------|---|

```

The frequency table in Output 3.4.1 displays the values of Eyes in order of descending frequency count. PROC FREQ computes the proportion of children in the first level displayed in the frequency table, Eyes = ‘brown’. Output 3.4.1 displays the binomial proportion confidence limits and test. The confidence limits are
90% confidence limits. If you do not specify the ALPHA= option, PROC FREQ computes 95% confidence limits by default. Because the value of Z is less than zero, PROC FREQ displays the a left-sided p-value (0.0019). This small p-value supports the alternative hypothesis that the true value of the proportion of children with brown eyes is less than 50%.

Output 3.4.2 displays the equivalence test results produced by the second TABLES statement. The null hypothesis proportion is 0.28 and the equivalence margins are –0.1 and 0.1, which yield equivalence limits of 0.18 and 0.38. PROC FREQ provides two one-sided tests (TOST) for equivalence. The small p-value indicates rejection of the null hypothesis in favor of the alternative that the proportion is equivalent to the null value.

**Output 3.4.2** Binomial Proportion for Hair Color

<table>
<thead>
<tr>
<th>Hair</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>fair</td>
<td>228</td>
<td>29.92</td>
<td>228</td>
<td>29.92</td>
</tr>
<tr>
<td>medium</td>
<td>217</td>
<td>28.48</td>
<td>445</td>
<td>58.40</td>
</tr>
<tr>
<td>dark</td>
<td>182</td>
<td>23.88</td>
<td>627</td>
<td>82.28</td>
</tr>
<tr>
<td>red</td>
<td>113</td>
<td>14.83</td>
<td>740</td>
<td>97.11</td>
</tr>
<tr>
<td>black</td>
<td>22</td>
<td>2.89</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Equivalence Analysis

\[
H_0: P - p_0 \leq \text{Lower Margin} \quad \text{or} \quad P - p_0 \geq \text{Upper Margin}
\]

\[
H_a: \text{Lower Margin} < P - p_0 < \text{Upper Margin}
\]

\[
p_0 = 0.28 \quad \text{Lower Margin} = -0.1 \quad \text{Upper Margin} = 0.1
\]

Proportion \quad ASE (Sample)

\[
0.2992 \quad 0.0166
\]

Two One-Sided Tests (TOST)

<table>
<thead>
<tr>
<th>Test</th>
<th>Z</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Margin</td>
<td>7.1865</td>
<td>Pr &gt; Z &lt; .0001</td>
</tr>
<tr>
<td>Upper Margin</td>
<td>-4.8701</td>
<td>Pr &lt; Z &lt; .0001</td>
</tr>
<tr>
<td>Overall</td>
<td>-5.0566</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Equivalence Limits \quad 90% Confidence Limits

\[
0.1800 \quad 0.3800 \quad 0.2719 \quad 0.3265
\]
Example 3.5: Analysis of a 2x2 Contingency Table

This example computes chi-square tests and Fisher’s exact test to compare the probability of coronary heart disease for two types of diet. It also estimates the relative risks and computes exact confidence limits for the odds ratio.

The data set FatComp contains hypothetical data for a case-control study of high fat diet and the risk of coronary heart disease. The data are recorded as cell counts, where the variable Count contains the frequencies for each exposure and response combination. The data set is sorted in descending order by the variables Exposure and Response, so that the first cell of the $2 \times 2$ table contains the frequency of positive exposure and positive response. The FORMAT procedure creates formats to identify the type of exposure and response with character values.

```plaintext
proc format;
   value ExpFmt 1='High Cholesterol Diet'
                0='Low Cholesterol Diet';
   value RspFmt 1='Yes'
                0='No';
run;

data FatComp;
   input Exposure Response Count;
   label Response='Heart Disease';
   datalines;
   0 0 6
   0 1 2
   1 0 4
   1 1 11
;
proc sort data=FatComp;
   by descending Exposure descending Response;
run;
```

In the following PROC FREQ statements, ORDER=DATA option orders the contingency table values by their order in the input data set. The TABLES statement requests a two-way table of Exposure by Response. The CHISQ option produces several chi-square tests, while the RELRISK option produces relative risk measures. The EXACT statement requests the exact Pearson chi-square test and exact confidence limits for the odds ratio.

```plaintext
proc freq data=FatComp order=data;
   format Exposure ExpFmt. Response RspFmt.;
   tables Exposure*Response / chisq relrisk;
   exact pchi or;
   weight Count;
   title 'Case-Control Study of High Fat/Cholesterol Diet';
run;
```
The contingency table in Output 3.5.1 displays the variable values so that the first table cell contains the frequency for the first cell in the data set (the frequency of positive exposure and positive response).

**Output 3.5.1** Contingency Table

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Response (Heart Disease)</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Cholesterol Diet</td>
<td>Yes</td>
<td>11</td>
<td>47.83</td>
<td>73.33</td>
<td>84.62</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>4</td>
<td>17.39</td>
<td>26.67</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>Low Cholesterol Diet</td>
<td>Yes</td>
<td>2</td>
<td>8.70</td>
<td>25.00</td>
<td>15.38</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>6</td>
<td>26.09</td>
<td>75.00</td>
<td>60.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13</td>
<td>56.52</td>
<td>56.52</td>
<td>100.00</td>
<td>23</td>
</tr>
</tbody>
</table>

Output 3.5.2 displays the chi-square statistics. Because the expected counts in some of the table cells are small, PROC FREQ gives a warning that the asymptotic chi-square tests might not be appropriate. In this case, the exact tests are appropriate. The alternative hypothesis for this analysis states that coronary heart disease is more likely to be associated with a high fat diet, so a one-sided test is desired. Fisher’s exact right-sided test analyzes whether the probability of heart disease in the high fat group exceeds the probability of heart disease in the low fat group; because this \( p \)-value is small, the alternative hypothesis is supported.

The odds ratio, displayed in Output 3.5.3, provides an estimate of the relative risk when an event is rare. This estimate indicates that the odds of heart disease is 8.25 times higher in the high fat diet group; however, the wide confidence limits indicate that this estimate has low precision.
Output 3.5.2  Chi-Square Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>4.9597</td>
<td>0.0259</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>5.0975</td>
<td>0.0240</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>3.1879</td>
<td>0.0742</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>4.7441</td>
<td>0.0294</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.4644</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.4212</td>
<td></td>
</tr>
<tr>
<td>Cramer's V</td>
<td></td>
<td>0.4644</td>
<td></td>
</tr>
</tbody>
</table>

WARNING: 50% of the cells have expected counts less than 5.  
(Asymptotic) Chi-Square may not be a valid test.

Pearson Chi-Square Test

Chi-Square  4.9597  
DF  1  
Asymptotic Pr > ChiSq  0.0259  
Exact  Pr >= ChiSq  0.0393

Fisher's Exact Test

Cell (1,1) Frequency (F)  11  
Left-sided Pr <= F  0.9967  
Right-sided Pr >= F  0.0367  
Table Probability (P)  0.0334  
Two-sided Pr <= P  0.0393

Output 3.5.3  Relative Risk

Odds Ratio and Relative Risks

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds Ratio</td>
<td>8.2500</td>
<td>1.1535 59.0029</td>
</tr>
<tr>
<td>Relative Risk (Column 1)</td>
<td>2.9333</td>
<td>0.8502 10.1204</td>
</tr>
<tr>
<td>Relative Risk (Column 2)</td>
<td>0.3556</td>
<td>0.1403 0.9009</td>
</tr>
</tbody>
</table>

Odds Ratio

Asymptotic Conf Limits
95% Lower Conf Limit  1.1535  
95% Upper Conf Limit  59.0029

Exact Conf Limits
95% Lower Conf Limit  0.8677  
95% Upper Conf Limit  105.5488
Example 3.6: Output Data Set of Chi-Square Statistics

This example uses the Color data from Example 3.1 to output the Pearson chi-square and the likelihood ratio chi-square statistics to a SAS data set. The following PROC FREQ statements create a two-way table of eye color versus hair color.

```sas
proc freq data=Color order=data;
  tables Eyes*Hair / expected cellchi2 norow nocol chisq;
  output out=ChiSqData n nmiss pchi lrchi;
  weight Count;
  title 'Chi-Square Tests for 3 by 5 Table of Eye and Hair Color';
run;

proc print data=ChiSqData noobs;
  title1 'Chi-Square Statistics for Eye and Hair Color';
  title2 'Output Data Set from the FREQ Procedure';
run;
```

The EXPECTED option displays expected cell frequencies in the crosstabulation table, and the CELLCHI2 option displays the cell contribution to the overall chi-square. The NOROW and NOCOL options suppress the display of row and column percents in the crosstabulation table. The CHISQ option produces chi-square tests.

The OUTPUT statement creates the ChiSqData output data set and specifies the statistics to include. The N option requests the number of nonmissing observations, the NMISS option stores the number of missing observations, and the PCHI and LRCHI options request Pearson and likelihood ratio chi-square statistics, respectively, together with their degrees of freedom and p-values.

The preceding statements produce Output 3.6.1 and Output 3.6.2. The contingency table in Output 3.6.1 displays eye and hair color in the order in which they appear in the Color data set. The Pearson chi-square statistic in Output 3.6.2 provides evidence of an association between eye and hair color (p=0.0073). The cell chi-square values show that most of the association is due to more green-eyed children with fair or red hair and fewer with dark or black hair. The opposite occurs with the brown-eyed children.

Output 3.6.3 displays the output data set created by the OUTPUT statement. It includes one observation that contains the sample size, the number of missing values, and the chi-square statistics and corresponding degrees of freedom and p-values as in Output 3.6.2.
### Output 3.6.1 Contingency Table

Chi-Square Tests for 3 by 5 Table of Eye and Hair Color

The FREQ Procedure

Table of Eyes by Hair

<table>
<thead>
<tr>
<th>Eyes(Eye Color)</th>
<th>Hair(Hair Color)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
</tr>
<tr>
<td>blue</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>66.425</td>
</tr>
<tr>
<td></td>
<td>0.0998</td>
</tr>
<tr>
<td></td>
<td>9.06</td>
</tr>
<tr>
<td>green</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>59.543</td>
</tr>
<tr>
<td></td>
<td>1.5019</td>
</tr>
<tr>
<td></td>
<td>9.06</td>
</tr>
<tr>
<td>brown</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>102.03</td>
</tr>
<tr>
<td></td>
<td>1.4187</td>
</tr>
<tr>
<td></td>
<td>11.81</td>
</tr>
<tr>
<td>Total</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>29.92</td>
</tr>
</tbody>
</table>

### Output 3.6.2 Chi-Square Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>8</td>
<td>20.9248</td>
<td>0.0073</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>8</td>
<td>25.9733</td>
<td>0.0011</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>3.7838</td>
<td>0.0518</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.1657</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.1635</td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.1172</td>
<td></td>
</tr>
</tbody>
</table>

### Output 3.6.3 Output Data Set

Chi-Square Statistics for Eye and Hair Color

Output Data Set from the FREQ Procedure

<table>
<thead>
<tr>
<th>N</th>
<th>NMISS</th>
<th><em>PCHI</em></th>
<th>DF_PCHI</th>
<th>P_PCHI</th>
<th><em>LRCHI</em></th>
<th>DF_LRCHI</th>
<th>P_LRCHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>762</td>
<td>0</td>
<td>20.9248</td>
<td>8</td>
<td>.007349898</td>
<td>25.9733</td>
<td>8</td>
<td>.001061424</td>
</tr>
</tbody>
</table>
Example 3.7: Cochran-Mantel-Haenszel Statistics

The data set Migraine contains hypothetical data for a clinical trial of migraine treatment. Subjects of both genders receive either a new drug therapy or a placebo. Their response to treatment is coded as ‘Better’ or ‘Same’. The data are recorded as cell counts, and the number of subjects for each treatment and response combination is recorded in the variable Count.

```r
data Migraine;
input Gender $ Treatment $ Response $ Count @@;
datalines;
  female Active Better 16 female Active Same 11
  female Placebo Better 5 female Placebo Same 20
  male Active Better 12 male Active Same 16
  male Placebo Better 7 male Placebo Same 19;
```

The following PROC FREQ statements create a multiway table stratified by Gender, where Treatment forms the rows and Response forms the columns. The RELRISK option in the TABLES statement requests the odds ratio and relative risks for the two-way tables of Treatment by Response. The PLOTS= option requests a relative risk plot, which shows the relative risk and its confidence limits for each level of Gender and overall. The CMH option requests Cochran-Mantel-Haenszel statistics for the multiway table. For this stratified 2 x 2 table, the CMH option also produces estimates of the common relative risk and the Breslow-Day test for homogeneity of the odds ratios. The NOPRINT option suppresses the display of the crosstabulation tables.

```r
ods graphics on;
proc freq data=Migraine;
tables Gender*Treatment*Response /
   relrisk plots(only)=relriskplot(stats) cmh noprint;
weight Count;
title 'Clinical Trial for Treatment of Migraine Headaches';
run;
ods graphics off;
```

Output 3.7.1 through Output 3.7.4 show the results of the analysis. The relative risk plot (Output 3.7.1) displays the relative risks and confidence limits for the two levels of Gender and for the overall (common) relative risk. Output 3.7.2 displays the CMH statistics. For a stratified 2 x 2 table, the three CMH statistics test the same hypothesis. The significant p-value (0.004) indicates that the association between treatment and response remains strong after adjusting for gender.

The CMH option also produces a table of overall relative risks, as shown in Output 3.7.3. Because this is a prospective study, the relative risk estimate assesses the effectiveness of the new drug; the “Cohort (Col1 Risk)” values are the appropriate estimates for the first column (the risk of improvement). The probability of migraine improvement with the new drug is just over two times the probability of improvement with the placebo.

The large p-value for the Breslow-Day test (0.2218) in Output 3.7.4 indicates no significant gender difference in the odds ratios.
Chapter 3: The FREQ Procedure

Output 3.7.1 Relative Risk Plot

Relative Risks with 95% Confidence Limits
Treatment by Response

- Female: 2.963 (1.274, 6.8913)
- Male: 1.5918 (0.7413, 3.418)
- Common: 2.1636 (1.2336, 3.7948)

Risks computed for column 1 (Response = Better)

Output 3.7.2 Cochran-Mantel-Haenszel Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
<tr>
<td>2</td>
<td>Row Mean Scores Differ</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
<tr>
<td>3</td>
<td>General Association</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
</tbody>
</table>
Example 3.8: Cochran-Armitage Trend Test

The data set Pain contains hypothetical data for a clinical trial of a drug therapy to control pain. The clinical trial investigates whether adverse responses increase with larger drug doses. Subjects receive either a placebo or one of four drug doses. An adverse response is recorded as Adverse='Yes'; otherwise, it is recorded as Adverse='No'. The number of subjects for each drug dose and response combination is contained in the variable Count.

data pain;
    input Dose Adverse $ Count @@;
datalines;
  0 No  26  0 Yes  6
  1 No  26  1 Yes  7
  2 No  23  2 Yes  9
  3 No  18  3 Yes 14
  4 No  9   4 Yes 23
;

The following PROC FREQ statements provide a trend analysis. The TABLES statement requests a table of Adverse by Dose. The MEASURES option produces measures of association, and the CL option produces confidence limits for these measures. The TREND option tests for a trend across the ordinal values of the variable Dose with the Cochran-Armitage test. The PLOTS= option requests a mosaic plot of Adverse by Dose.

Output 3.7.3 CMH Option: Common Relative Risks

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Method</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds Ratio</td>
<td>Mantel-Haenszel</td>
<td>3.3132</td>
<td>1.4456</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>3.2941</td>
<td>1.4182</td>
</tr>
<tr>
<td>Relative Risk (Column 1)</td>
<td>Mantel-Haenszel</td>
<td>2.1636</td>
<td>1.2336</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>2.1059</td>
<td>1.1951</td>
</tr>
<tr>
<td>Relative Risk (Column 2)</td>
<td>Mantel-Haenszel</td>
<td>0.6420</td>
<td>0.4705</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>0.6613</td>
<td>0.4852</td>
</tr>
</tbody>
</table>

Output 3.7.4 CMH Option: Breslow-Day Test

<table>
<thead>
<tr>
<th>Breslow-Day Test for Homogeneity of the Odds Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square: 1.4929</td>
</tr>
<tr>
<td>DF: 1</td>
</tr>
<tr>
<td>Pr &gt; ChiSq: 0.2218</td>
</tr>
</tbody>
</table>

Example 3.8: Cochran-Armitage Trend Test

The data set Pain contains hypothetical data for a clinical trial of a drug therapy to control pain. The clinical trial investigates whether adverse responses increase with larger drug doses. Subjects receive either a placebo or one of four drug doses. An adverse response is recorded as Adverse='Yes'; otherwise, it is recorded as Adverse='No'. The number of subjects for each drug dose and response combination is contained in the variable Count.

data pain;
    input Dose Adverse $ Count @@;
datalines;
  0 No  26  0 Yes  6
  1 No  26  1 Yes  7
  2 No  23  2 Yes  9
  3 No  18  3 Yes 14
  4 No  9   4 Yes 23
;

The following PROC FREQ statements provide a trend analysis. The TABLES statement requests a table of Adverse by Dose. The MEASURES option produces measures of association, and the CL option produces confidence limits for these measures. The TREND option tests for a trend across the ordinal values of the variable Dose with the Cochran-Armitage test. The PLOTS= option requests a mosaic plot of Adverse by Dose.
The EXACT statement produces exact \( p \)-values for this test, and the MAXTIME= option terminates the exact computations if they do not complete within 60 seconds. The TEST statement computes an asymptotic test for Somers’ \( D(R|C) \).

```plaintext
ods graphics on;
proc freq data=Pain;
   tables Adverse*Dose / trend measures cl
   plots=mosaicplot;
   test smdrc;
   exact trend / maxtime=60;
   weight Count;
   title 'Clinical Trial for Treatment of Pain';
run;
ods graphics off;
```

Output 3.8.1 through Output 3.8.4 display the results of the analysis. The “Col Pct” values in Output 3.8.1 show the expected increasing trend in the proportion of adverse effects with the increasing dosage (from 18.75% to 71.88%). The corresponding mosaic plot (Output 3.8.2) also shows this increasing trend.

**Output 3.8.1** Contingency Table

```
Clinical Trial for Treatment of Pain

The FREQ Procedure

Table of Adverse by Dose

<table>
<thead>
<tr>
<th>Adverse</th>
<th>Dose</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>No</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>23</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
</tr>
</tbody>
</table>
```
Output 3.8.3 displays the measures of association produced by the MEASURES option. Somers’ $D(R|C)$ measures the association treating the row variable (Adverse) as the response and the column variable (Dose) as a predictor. Because the asymptotic 95% confidence limits do not contain zero, this indicates a strong positive association. Similarly, the Pearson and Spearman correlation coefficients show evidence of a strong positive association, as hypothesized.

The Cochran-Armitage test (Output 3.8.4) supports the trend hypothesis. The small left-sided $p$-values for the Cochran-Armitage test indicate that the probability of the Row 1 level (Adverse=‘No’) decreases as Dose increases or, equivalently, that the probability of the Row 2 level (Adverse=‘Yes’) increases as Dose increases. The two-sided $p$-value tests against either an increasing or decreasing alternative. This is an appropriate hypothesis when you want to determine whether the drug has progressive effects on the probability of adverse effects but the direction is unknown.
Chapter 3: The FREQ Procedure

Output 3.8.3 Measures of Association

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.5313</td>
<td>0.0935</td>
<td>0.3480 0.7146</td>
</tr>
<tr>
<td>Kendall's Tau-b</td>
<td>0.3373</td>
<td>0.0642</td>
<td>0.2114 0.4631</td>
</tr>
<tr>
<td>Stuart's Tau-c</td>
<td>0.4111</td>
<td>0.0798</td>
<td>0.2547 0.5675</td>
</tr>
<tr>
<td>Somers' D C</td>
<td>R</td>
<td>0.4427</td>
<td>0.0837</td>
</tr>
<tr>
<td>Somers' D R</td>
<td>C</td>
<td>0.2569</td>
<td>0.0499</td>
</tr>
<tr>
<td>Spearman Correlation</td>
<td>0.3776</td>
<td>0.0714</td>
<td>0.2378 0.5175</td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>0.3771</td>
<td>0.0718</td>
<td>0.2363 0.5178</td>
</tr>
<tr>
<td>Lambda Asymmetric C</td>
<td>R</td>
<td>0.1250</td>
<td>0.0662</td>
</tr>
<tr>
<td>Lambda Asymmetric R</td>
<td>C</td>
<td>0.2373</td>
<td>0.0837</td>
</tr>
<tr>
<td>Lambda Symmetric</td>
<td>0.1604</td>
<td>0.0621</td>
<td>0.0388 0.2821</td>
</tr>
<tr>
<td>Uncertainty Coefficient C</td>
<td>R</td>
<td>0.0515</td>
<td>0.0191</td>
</tr>
<tr>
<td>Uncertainty Coefficient R</td>
<td>C</td>
<td>0.1261</td>
<td>0.0467</td>
</tr>
<tr>
<td>Uncertainty Coefficient Symmetric</td>
<td>0.0731</td>
<td>0.0271</td>
<td>0.0199 0.1262</td>
</tr>
</tbody>
</table>

Somers' D R|C

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Somers' D R</td>
<td>C</td>
<td>0.2569</td>
<td>0.0499</td>
</tr>
<tr>
<td>ASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% Lower Conf Limit</td>
<td></td>
<td>0.1592</td>
<td></td>
</tr>
<tr>
<td>95% Upper Conf Limit</td>
<td>0.3547</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test of H0: Somers' D R|C = 0

| Test Statistic | Value  | Pr > | |Z| |
|----------------|--------|------|------|
| ASE under H0   | 0.0499 | <.0001|
| Z               | 5.1511 |      |
| One-sided Pr > | Z     | <.0001|
| Two-sided Pr > | |Z| | <.0001|

Output 3.8.4 Trend Test

Cochran-Armitage Trend Test

<table>
<thead>
<tr>
<th>Statistic (Z)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.7918</td>
<td></td>
</tr>
</tbody>
</table>

Asymptotic Test

| Test Statistic | Value  | Pr > | |Z| |
|----------------|--------|------|------|
| One-sided Pr < | Z     | <.0001|
| Two-sided Pr > | |Z| | <.0001|

Exact Test

| Test Statistic | Value  | Pr > | |Z| |
|----------------|--------|------|------|
| One-sided Pr <= | Z | <.0001|
| Two-sided Pr >= | |Z| | <.0001|
Example 3.9: Friedman’s Chi-Square Test

Friedman’s test is a nonparametric test for treatment differences in a randomized complete block design. Each block of the design might be a subject or a homogeneous group of subjects. If blocks are groups of subjects, the number of subjects in each block must equal the number of treatments. Treatments are randomly assigned to subjects within each block. If there is one subject per block, then the subjects are repeatedly measured once under each treatment. The order of treatments is randomized for each subject.

In this setting, Friedman’s test is identical to the ANOVA (row means scores) CMH statistic when the analysis uses rank scores (SCORES=RANK). The three-way table uses subject (or subject group) as the stratifying variable, treatment as the row variable, and response as the column variable. PROC FREQ handles ties by assigning midranks to tied response values. If there are multiple subjects per treatment in each block, the ANOVA CMH statistic is a generalization of Friedman’s test.

The data set Hypnosis contains data from a study investigating whether hypnosis has the same effect on skin potential (measured in millivolts) for four emotions (Lehmann and D’Abrera 2006, p. 264). Eight subjects are asked to display fear, joy, sadness, and calmness under hypnosis. The data are recorded as one observation per subject for each emotion.

```sas
data Hypnosis;
  length Emotion $ 10;
  input Subject Emotion $ SkinResponse @@;
datalines;
1 fear 23.1 1 joy 22.7 1 sadness 22.5 1 calmness 22.6
2 fear 57.6 2 joy 53.2 2 sadness 53.7 2 calmness 53.1
3 fear 10.5 3 joy 9.7 3 sadness 10.8 3 calmness 8.3
4 fear 23.6 4 joy 19.6 4 sadness 21.1 4 calmness 21.6
5 fear 11.9 5 joy 13.8 5 sadness 13.7 5 calmness 13.3
6 fear 54.6 6 joy 47.1 6 sadness 39.2 6 calmness 37.0
7 fear 21.0 7 joy 13.6 7 sadness 13.7 7 calmness 14.8
8 fear 20.3 8 joy 23.6 8 sadness 16.3 8 calmness 14.8
;
```

In the following PROC FREQ statements, the TABLES statement creates a three-way table stratified by Subject and a two-way table; the variables Emotion and SkinResponse form the rows and columns of each table. The CMH2 option produces the first two Cochran-Mantel-Haenszel statistics, the option SCORES=RANK specifies that rank scores are used to compute these statistics, and the NOPRINT option suppresses the contingency tables. These statements produce Output 3.9.1 and Output 3.9.2.

```sas
proc freq data=Hypnosis;
  tables Subject*Emotion*SkinResponse / cmh2 scores=rank noprint;
run;

proc freq data=Hypnosis;
  tables Emotion*SkinResponse / cmh2 scores=rank noprint;
run;
```

Because the CMH statistics in Output 3.9.1 are based on rank scores, the Row Mean Scores Differ statistic is identical to Friedman’s chi-square ($Q = 6.45$). The $p$-value of 0.0917 indicates that differences in skin potential response for different emotions are significant at the 10% level but not at the 5% level.
When you do not stratify by subject, the Row Mean Scores Differ CMH statistic is identical to a Kruskal-Wallis test and is not significant (p = 0.9038 in Output 3.9.2). Thus, adjusting for subject is critical to reducing the background variation due to subject differences.

**Output 3.9.1** CMH Statistics: Stratifying by Subject

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Nonzero Correlation</td>
<td>1</td>
<td>1</td>
<td>0.2400</td>
<td>0.6242</td>
</tr>
<tr>
<td>2: Row Mean Scores Differ</td>
<td>3</td>
<td>3</td>
<td>6.4500</td>
<td>0.0917</td>
</tr>
</tbody>
</table>

**Output 3.9.2** CMH Statistics: No Stratification

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Nonzero Correlation</td>
<td>1</td>
<td>1</td>
<td>0.0001</td>
<td>0.9933</td>
</tr>
<tr>
<td>2: Row Mean Scores Differ</td>
<td>3</td>
<td>3</td>
<td>0.5678</td>
<td>0.9038</td>
</tr>
</tbody>
</table>

**Example 3.10: Cochran’s Q Test**

When a binary response is measured several times or under different conditions, Cochran’s Q tests that the marginal probability of a positive response is unchanged across the times or conditions. When there are more than two response categories, you can use the CATMOD procedure to fit a repeated-measures model.

The data set Drugs contains data for a study of three drugs to treat a chronic disease (Agresti 2002). Forty-six subjects receive drugs A, B, and C. The response to each drug is either favorable (‘F’) or unfavorable (‘U’).

```plaintext
proc format;
  value $ResponseFmt 'F'='Favorable'
                  'U'='Unfavorable';
run;
```
data drugs;
  input Drug_A $ Drug_B $ Drug_C $ Count @@;
datalines;
  F F F 6  U F F 2
  F F U 16  U F U 4
  F U F 2  U U F 6
  F U U 4  U U U 6
;

The following statements create one-way frequency tables of the responses to each drug. The AGREE option produces Cochran’s Q and other measures of agreement for the three-way table. These statements produce Output 3.10.1 through Output 3.10.5.

proc freq data=Drugs;
  tables Drug_A Drug_B Drug_C / nocum;
  tables Drug_A*Drug_B*Drug_C / agree noprint;
  format Drug_A Drug_B Drug_C $ResponseFmt.;
  weight Count;
  title 'Study of Three Drug Treatments for a Chronic Disease';
run;

The one-way frequency tables in Output 3.10.1 provide the marginal response for each drug. For drugs A and B, 61% of the subjects reported a favorable response while 35% of the subjects reported a favorable response to drug C. Output 3.10.2 and Output 3.10.3 display measures of agreement for the ‘Favorable’ and ‘Unfavorable’ levels of drug A, respectively. McNemar’s test shows a strong discordance between drugs B and C when the response to drug A is favorable.

**Output 3.10.1** One-Way Frequency Tables

<table>
<thead>
<tr>
<th>Drug_A</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>28</td>
<td>60.87</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>18</td>
<td>39.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug_B</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>28</td>
<td>60.87</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>18</td>
<td>39.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug_C</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>16</td>
<td>34.78</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>30</td>
<td>65.22</td>
</tr>
</tbody>
</table>
Output 3.10.2 Measures of Agreement for Drug A Favorable

McNemar's Test

<table>
<thead>
<tr>
<th>Statistic (S)</th>
<th>10.8889</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>1</td>
</tr>
<tr>
<td>Pr &gt; S</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Simple Kappa Coefficient

<table>
<thead>
<tr>
<th>Kappa</th>
<th>-0.0328</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>0.1167</td>
</tr>
<tr>
<td>95% Lower Conf Limit</td>
<td>-0.2615</td>
</tr>
<tr>
<td>95% Upper Conf Limit</td>
<td>0.1960</td>
</tr>
</tbody>
</table>

Output 3.10.3 Measures of Agreement for Drug A Unfavorable

McNemar's Test

<table>
<thead>
<tr>
<th>Statistic (S)</th>
<th>0.4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>1</td>
</tr>
<tr>
<td>Pr &gt; S</td>
<td>0.5271</td>
</tr>
</tbody>
</table>

Simple Kappa Coefficient

<table>
<thead>
<tr>
<th>Kappa</th>
<th>-0.1538</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>0.2230</td>
</tr>
<tr>
<td>95% Lower Conf Limit</td>
<td>-0.5909</td>
</tr>
<tr>
<td>95% Upper Conf Limit</td>
<td>0.2832</td>
</tr>
</tbody>
</table>

Output 3.10.4 displays the overall kappa coefficient. The small negative value of kappa indicates no agreement between drug B response and drug C response.

Output 3.10.4 Overall Measures of Agreement

Overall Kappa Coefficient

<table>
<thead>
<tr>
<th>Kappa</th>
<th>-0.0588</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>0.1034</td>
</tr>
<tr>
<td>95% Lower Conf Limit</td>
<td>-0.2615</td>
</tr>
<tr>
<td>95% Upper Conf Limit</td>
<td>0.1439</td>
</tr>
</tbody>
</table>

Test for Equal Kappa Coefficients

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>0.2314</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>1</td>
</tr>
<tr>
<td>Pr &gt; ChiSq</td>
<td>0.6305</td>
</tr>
</tbody>
</table>
Cochran’s $Q$ is statistically significant ($p=0.0145$ in **Output 3.10.5**), which leads to rejection of the hypothesis that the probability of favorable response is the same for the three drugs.

**Output 3.10.5** Cochran’s $Q$ Test

<table>
<thead>
<tr>
<th>Cochran’s $Q$, for Drug_A by Drug_B by Drug_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic (Q)</td>
</tr>
<tr>
<td>8.4706</td>
</tr>
<tr>
<td>DF</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>$Pr &gt; Q$</td>
</tr>
<tr>
<td>0.0145</td>
</tr>
</tbody>
</table>

---

**References**


Clopper, C. J. and Pearson, E. S. (1934), “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial,” Biometrika, 26, 404–413.


# Chapter 4
The UNIVARIATE Procedure

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview: UNIVARIATE Procedure</td>
<td>265</td>
</tr>
<tr>
<td>Getting Started: UNIVARIATE Procedure</td>
<td>266</td>
</tr>
<tr>
<td>Capabilities of PROC UNIVARIATE</td>
<td>266</td>
</tr>
<tr>
<td>Summarizing a Data Distribution</td>
<td>267</td>
</tr>
<tr>
<td>Exploring a Data Distribution</td>
<td>268</td>
</tr>
<tr>
<td>Modeling a Data Distribution</td>
<td>271</td>
</tr>
<tr>
<td>Syntax: UNIVARIATE Procedure</td>
<td>274</td>
</tr>
<tr>
<td>PROC UNIVARIATE Statement</td>
<td>274</td>
</tr>
<tr>
<td>BY Statement</td>
<td>282</td>
</tr>
<tr>
<td>CDFPLOT Statement</td>
<td>282</td>
</tr>
<tr>
<td>CLASS Statement</td>
<td>294</td>
</tr>
<tr>
<td>FREQ Statement</td>
<td>296</td>
</tr>
<tr>
<td>HISTOGRAM Statement</td>
<td>296</td>
</tr>
<tr>
<td>ID Statement</td>
<td>316</td>
</tr>
<tr>
<td>INSET Statement</td>
<td>316</td>
</tr>
<tr>
<td>OUTPUT Statement</td>
<td>326</td>
</tr>
<tr>
<td>PPLOT Statement</td>
<td>332</td>
</tr>
<tr>
<td>PROBLOT Statement</td>
<td>345</td>
</tr>
<tr>
<td>QQPLOT Statement</td>
<td>356</td>
</tr>
<tr>
<td>VAR Statement</td>
<td>367</td>
</tr>
<tr>
<td>WEIGHT Statement</td>
<td>367</td>
</tr>
<tr>
<td>Dictionary of Common Options</td>
<td>369</td>
</tr>
<tr>
<td>Details: UNIVARIATE Procedure</td>
<td>377</td>
</tr>
<tr>
<td>Missing Values</td>
<td>377</td>
</tr>
<tr>
<td>Rounding</td>
<td>378</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>379</td>
</tr>
<tr>
<td>Calculating the Mode</td>
<td>382</td>
</tr>
<tr>
<td>Calculating Percentiles</td>
<td>382</td>
</tr>
<tr>
<td>Tests for Location</td>
<td>384</td>
</tr>
<tr>
<td>Confidence Limits for Parameters of the Normal Distribution</td>
<td>386</td>
</tr>
<tr>
<td>Robust Estimators</td>
<td>387</td>
</tr>
<tr>
<td>Creating Line Printer Plots</td>
<td>389</td>
</tr>
<tr>
<td>Creating High-Resolution Graphics</td>
<td>391</td>
</tr>
<tr>
<td>Using the CLASS Statement to Create Comparative Plots</td>
<td>396</td>
</tr>
<tr>
<td>Positioning Insets</td>
<td>397</td>
</tr>
<tr>
<td>Formulas for Fitted Continuous Distributions</td>
<td>401</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Goodness-of-Fit Tests</td>
<td>413</td>
</tr>
<tr>
<td>Kernel Density Estimates</td>
<td>418</td>
</tr>
<tr>
<td>Construction of Quantile-Quantile and Probability Plots</td>
<td>420</td>
</tr>
<tr>
<td>Interpretation of Quantile-Quantile and Probability Plots</td>
<td>420</td>
</tr>
<tr>
<td>Distributions for Probability and Q-Q Plots</td>
<td>421</td>
</tr>
<tr>
<td>Estimating Shape Parameters Using Q-Q Plots</td>
<td>427</td>
</tr>
<tr>
<td>Estimating Location and Scale Parameters Using Q-Q Plots</td>
<td>427</td>
</tr>
<tr>
<td>Estimating Percentiles Using Q-Q Plots</td>
<td>428</td>
</tr>
<tr>
<td>Input Data Sets</td>
<td>429</td>
</tr>
<tr>
<td>OUT= Output Data Set in the OUTPUT Statement</td>
<td>429</td>
</tr>
<tr>
<td>OUTHISTOGRAM= Output Data Set</td>
<td>431</td>
</tr>
<tr>
<td>OUTKERNEL= Output Data Set</td>
<td>432</td>
</tr>
<tr>
<td>OUTTABLE= Output Data Set</td>
<td>433</td>
</tr>
<tr>
<td>Tables for Summary Statistics</td>
<td>435</td>
</tr>
<tr>
<td>ODS Table Names</td>
<td>436</td>
</tr>
<tr>
<td>ODS Tables for Fitted Distributions</td>
<td>437</td>
</tr>
<tr>
<td>ODS Graphics</td>
<td>437</td>
</tr>
<tr>
<td>Computational Resources</td>
<td>438</td>
</tr>
<tr>
<td>Examples: UNIVARIATE Procedure</td>
<td>439</td>
</tr>
<tr>
<td>Example 4.1: Computing Descriptive Statistics for Multiple Variables</td>
<td>439</td>
</tr>
<tr>
<td>Example 4.2: Calculating Modes</td>
<td>441</td>
</tr>
<tr>
<td>Example 4.3: Identifying Extreme Observations and Extreme Values</td>
<td>443</td>
</tr>
<tr>
<td>Example 4.4: Creating a Frequency Table</td>
<td>445</td>
</tr>
<tr>
<td>Example 4.5: Creating Plots for Line Printer Output</td>
<td>447</td>
</tr>
<tr>
<td>Example 4.6: Analyzing a Data Set With a FREQ Variable</td>
<td>452</td>
</tr>
<tr>
<td>Example 4.7: Saving Summary Statistics in an OUT= Output Data Set</td>
<td>453</td>
</tr>
<tr>
<td>Example 4.8: Saving Percentiles in an Output Data Set</td>
<td>454</td>
</tr>
<tr>
<td>Example 4.9: Computing Confidence Limits for the Mean, Standard Deviation, and Variance</td>
<td>455</td>
</tr>
<tr>
<td>Example 4.10: Computing Confidence Limits for Quantiles and Percentiles</td>
<td>457</td>
</tr>
<tr>
<td>Example 4.11: Computing Robust Estimates</td>
<td>460</td>
</tr>
<tr>
<td>Example 4.12: Testing for Location</td>
<td>462</td>
</tr>
<tr>
<td>Example 4.13: Performing a Sign Test Using Paired Data</td>
<td>463</td>
</tr>
<tr>
<td>Example 4.14: Creating a Histogram</td>
<td>464</td>
</tr>
<tr>
<td>Example 4.15: Creating a One-Way Comparative Histogram</td>
<td>466</td>
</tr>
<tr>
<td>Example 4.16: Creating a Two-Way Comparative Histogram</td>
<td>469</td>
</tr>
<tr>
<td>Example 4.17: Adding Insets with Descriptive Statistics</td>
<td>471</td>
</tr>
<tr>
<td>Example 4.18: Binning a Histogram</td>
<td>473</td>
</tr>
<tr>
<td>Example 4.19: Adding a Normal Curve to a Histogram</td>
<td>476</td>
</tr>
<tr>
<td>Example 4.20: Adding Fitted Normal Curves to a Comparative Histogram</td>
<td>479</td>
</tr>
<tr>
<td>Example 4.21: Fitting a Beta Curve</td>
<td>481</td>
</tr>
<tr>
<td>Example 4.22: Fitting Lognormal, Weibull, and Gamma Curves</td>
<td>483</td>
</tr>
<tr>
<td>Example 4.23: Computing Kernel Density Estimates</td>
<td>488</td>
</tr>
<tr>
<td>Example 4.24: Fitting a Three-Parameter Lognormal Curve</td>
<td>489</td>
</tr>
</tbody>
</table>
Overview: UNIVARIATE Procedure

The UNIVARIATE procedure provides the following:

- descriptive statistics based on moments (including skewness and kurtosis), quantiles or percentiles (such as the median), frequency tables, and extreme values
- histograms that optionally can be fitted with probability density curves for various distributions and with kernel density estimates
- cumulative distribution function plots (cdf plots). Optionally, these can be superimposed with probability distribution curves for various distributions.
- quantile-quantile plots (Q-Q plots), probability plots, and probability-probability plots (P-P plots). These plots facilitate the comparison of a data distribution with various theoretical distributions.
- goodness-of-fit tests for a variety of distributions including the normal
- the ability to inset summary statistics on plots
- the ability to analyze data sets with a frequency variable
- the ability to create output data sets containing summary statistics, histogram intervals, and parameters of fitted curves

You can use the PROC UNIVARIATE statement, together with the VAR statement, to compute summary statistics. See the section “Getting Started: UNIVARIATE Procedure” on page 266 for introductory examples. In addition, you can use the following statements to request plots:

- the CDFPLOT statement for creating cdf plots
- the HISTOGRAM statement for creating histograms
• the `PPLOT` statement for creating P-P plots
• the `PROBPLOT` statement for creating probability plots
• the `QQPLOT` statement for creating Q-Q plots
• the `CLASS` statement together with any of these plot statements for creating comparative plots
• the `INSET` statement with any of the plot statements for enhancing the plot with an inset table of summary statistics

The UNIVARIATE procedure produces two kinds of graphical output:

• ODS Statistical Graphics output, which is produced when ODS Graphics is enabled prior to your procedure statements
• traditional graphics, which are produced when ODS Graphics is not enabled

See the section “Creating High-Resolution Graphics” on page 391 for more information about producing traditional graphics and ODS Graphics output.

---

**Getting Started: UNIVARIATE Procedure**

The following examples demonstrate how you can use the UNIVARIATE procedure to analyze the distributions of variables through the use of descriptive statistical measures and graphical displays, such as histograms.

---

**Capabilities of PROC UNIVARIATE**

The UNIVARIATE procedure provides a variety of descriptive measures, graphical displays, and statistical methods, which you can use to summarize, visualize, analyze, and model the statistical distributions of numeric variables. These tools are appropriate for a broad range of tasks and applications:

• Exploring the distributions of the variables in a data set is an important preliminary step in data analysis, data warehousing, and data mining. With the UNIVARIATE procedure you can use tables and graphical displays, such as histograms and nonparametric density estimates, to find key features of distributions, identify outliers and extreme observations, determine the need for data transformations, and compare distributions.

• Modeling the distributions of data and validating distributional assumptions are basic steps in statistical analysis. You can use the UNIVARIATE procedure to fit parametric distributions (beta, exponential, gamma, Gumbel, inverse Gaussian, lognormal, normal, generalized Pareto, power function, Rayleigh, Johnson $S_B$, Johnson $S_U$, and Weibull) and to compute probabilities and percentiles from these models. You can assess goodness of fit with hypothesis tests and with graphical displays such as probability plots and quantile-quantile plots. You can also use the UNIVARIATE procedure to validate
distributional assumptions for other types of statistical analysis. When standard assumptions are not met, you can use the UNIVARIATE procedure to perform nonparametric tests and compute robust estimates of location and scale.

- Summarizing the distribution of the data is often helpful for creating effective statistical reports and presentations. You can use the UNIVARIATE procedure to create tables of summary measures, such as means and percentiles, together with graphical displays, such as histograms and comparative histograms, which facilitate the interpretation of the report.

The following examples illustrate a few of the tasks that you can carry out with the UNIVARIATE procedure.

---

**Summarizing a Data Distribution**

Figure 4.1 shows a table of basic summary measures and a table of extreme observations for the loan-to-value ratios of 5,840 home mortgages. The ratios are saved as values of the variable LoanToValueRatio in a data set named HomeLoans. The following statements request a univariate analysis:

```plaintext
ods select BasicMeasures ExtremeObs;
proc univariate data=HomeLoans;
    var LoanToValueRatio;
run;
```

The ODS SELECT statement restricts the default output to the tables for basic statistical measures and extreme observations.

**Figure 4.1 Basic Measures and Extreme Observations**

<table>
<thead>
<tr>
<th>The UNIVARIATE Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable: LoanToValueRatio (Loan to Value Ratio)</td>
</tr>
</tbody>
</table>

**Basic Statistical Measures**

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.292512</td>
</tr>
<tr>
<td>Median</td>
<td>0.248050</td>
</tr>
<tr>
<td>Mode</td>
<td>0.250000</td>
</tr>
</tbody>
</table>

| Std Deviation | 0.16476 |
| Variance      | 0.02715 |
| Range         | 1.24780 |
| Interquartile Range | 0.16419 |

**Extreme Observations**

<table>
<thead>
<tr>
<th>Value</th>
<th>Obs</th>
<th>Value</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0651786</td>
<td>1</td>
<td>1.13976</td>
<td>5776</td>
</tr>
<tr>
<td>0.0690157</td>
<td>3</td>
<td>1.14209</td>
<td>5791</td>
</tr>
<tr>
<td>0.0699755</td>
<td>59</td>
<td>1.14286</td>
<td>5801</td>
</tr>
<tr>
<td>0.0702412</td>
<td>84</td>
<td>1.17090</td>
<td>5799</td>
</tr>
<tr>
<td>0.0704787</td>
<td>4</td>
<td>1.31298</td>
<td>5811</td>
</tr>
</tbody>
</table>
The tables in Figure 4.1 show, in particular, that the average ratio is 0.2925 and the minimum and maximum ratios are 0.06518 and 1.1398, respectively.

**Exploring a Data Distribution**

Figure 4.2 shows a histogram of the loan-to-value ratios. The histogram reveals features of the ratio distribution, such as its skewness and the peak at 0.175, which are not evident from the tables in the previous example. The following statements create the histogram:

```
ods graphics off;
title 'Home Loan Analysis';
proc univariate data=HomeLoans noprint;
    histogram LoanToValueRatio;
    inset n = 'Number of Homes' / position=ne;
run;
```

By default, PROC UNIVARIATE produces traditional graphics output, and the basic appearance of the histogram is determined by the prevailing ODS style. The NOPRINT option suppresses the display of summary statistics. The INSET statement inserts the total number of analyzed home loans in the upper right (northeast) corner of the plot.
The data set `HomeLoans` contains a variable named `LoanType` that classifies the loans into two types: Gold and Platinum. It is useful to compare the distributions of `LoanToValueRatio` for the two types. The following statements request quantiles for each distribution and a comparative histogram, which are shown in Figure 4.3 and Figure 4.4.

```sas
title 'Comparison of Loan Types';
options nogstyle;
ods select Quantiles MyHist;
proc univariate data=HomeLoans;
  var LoanToValueRatio;
  class LoanType;
  histogram LoanToValueRatio / kernel(color=red)
               cfill=ltgray
               name='MyHist';
  inset n='Number of Homes' median='Median Ratio' (5.3) / position=ne;
  label LoanType = 'Type of Loan';
run;
options gstyle;
```

The ODS SELECT statement restricts the default output to the tables of quantiles and the graph produced by the `HISTOGRAM` statement, which is identified by the value specified by the `NAME=` option. The `CLASS`
statement specifies LoanType as a classification variable for the quantile computations and comparative histogram. The KERNEL option adds a smooth nonparametric estimate of the ratio density to each histogram. The INSET statement specifies summary statistics to be displayed directly in the graph.

The NOGSTYLE system option specifies that the ODS style not influence the appearance of the histogram. Instead, the CFILL= option determines the color of the histogram bars and the COLOR= option specifies the color of the kernel density curve.

**Figure 4.3** Quantiles for Loan-to-Value Ratio

<table>
<thead>
<tr>
<th></th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>1.0617647</td>
</tr>
<tr>
<td>99%</td>
<td>0.8974576</td>
</tr>
<tr>
<td>95%</td>
<td>0.6385908</td>
</tr>
<tr>
<td>90%</td>
<td>0.4471369</td>
</tr>
<tr>
<td>75% Q3</td>
<td>0.2985099</td>
</tr>
<tr>
<td>50% Median</td>
<td>0.2217033</td>
</tr>
<tr>
<td>25% Q1</td>
<td>0.1734568</td>
</tr>
<tr>
<td>10%</td>
<td>0.1411130</td>
</tr>
<tr>
<td>5%</td>
<td>0.1213079</td>
</tr>
<tr>
<td>1%</td>
<td>0.0942167</td>
</tr>
<tr>
<td>0% Min</td>
<td>0.0651786</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>1.312981</td>
</tr>
<tr>
<td>99%</td>
<td>1.050000</td>
</tr>
<tr>
<td>95%</td>
<td>0.691803</td>
</tr>
<tr>
<td>90%</td>
<td>0.549273</td>
</tr>
<tr>
<td>75% Q3</td>
<td>0.430160</td>
</tr>
<tr>
<td>50% Median</td>
<td>0.366168</td>
</tr>
<tr>
<td>25% Q1</td>
<td>0.314452</td>
</tr>
<tr>
<td>10%</td>
<td>0.273670</td>
</tr>
<tr>
<td>5%</td>
<td>0.253124</td>
</tr>
<tr>
<td>1%</td>
<td>0.231114</td>
</tr>
<tr>
<td>0% Min</td>
<td>0.215504</td>
</tr>
</tbody>
</table>
The output in Figure 4.3 shows that the median ratio for Platinum loans (0.366) is greater than the median ratio for Gold loans (0.222). The comparative histogram in Figure 4.4 enables you to compare the two distributions more easily. It shows that the ratio distributions are similar except for a shift of about 0.14.

Figure 4.4 Comparative Histogram for Loan-to-Value Ratio

A sample program for this example, *univar1.sas*, is available in the SAS Sample Library for Base SAS software.

---

Modeling a Data Distribution

In addition to summarizing a data distribution as in the preceding example, you can use PROC UNIVARIATE to statistically model a distribution based on a random sample of data. The following statements create a data set named Aircraft that contains the measurements of a position deviation for a sample of 30 aircraft components.

```sas
data Aircraft;
  input Deviation @@;
  label Deviation = 'Position Deviation';
datalines;
```
Chapter 4: The UNIVARIATE Procedure

An initial question in the analysis is whether the measurement distribution is normal. The following statements request a table of moments, the tests for normality, and a normal probability plot, which are shown in Figure 4.5 and Figure 4.6:

```
title 'Position Deviation Analysis';
odds graphics on;
odds select Moments TestsForNormality ProbPlot;
proc univariate data=Aircraft normaltest;
  var Deviation;
  probplot Deviation / normal (mu=est sigma=est)
    square;
  label Deviation = 'Position Deviation';
  inset mean std / format=6.4;
run;
```

When ODS Graphics is enabled, the procedure produces ODS Graphics output rather than traditional graphics. (See the section “Alternatives for Producing Graphics” on page 392 for information about traditional graphics and ODS Graphics.) The INSET statement displays the sample mean and standard deviation on the probability plot.

**Figure 4.5** Moments and Tests for Normality

<table>
<thead>
<tr>
<th>Moments</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.0053067</td>
<td>-0.1592</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>0.00254362</td>
<td>Variance 6.47002E-6</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.2562507</td>
<td>Kurtosis 0.69790426</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>0.00103245</td>
<td>Corrected SS 0.00018763</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>-47.932613</td>
<td>Std Error Mean 0.0004644</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tests for Normality</th>
<th>--Statistic--</th>
<th>------p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.845364</td>
<td>Pr &lt; W 0.0005</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.208921</td>
<td>Pr &gt; D &lt;0.0100</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.329274</td>
<td>Pr &gt; W-Sq &lt;0.0050</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 1.784881</td>
<td>Pr &gt; A-Sq &lt;0.0050</td>
</tr>
</tbody>
</table>
All four goodness-of-fit tests in Figure 4.5 reject the hypothesis that the measurements are normally distributed.

Figure 4.6 shows a normal probability plot for the measurements. A linear pattern of points following the diagonal reference line would indicate that the measurements are normally distributed. Instead, the curved point pattern suggests that a skewed distribution, such as the lognormal, is more appropriate than the normal distribution.

A lognormal distribution for Deviation is fitted in Example 4.26.

A sample program for this example, univar2.sas, is available in the SAS Sample Library for Base SAS software.
## Syntax: UNIVARIATE Procedure

```
PROC UNIVARIATE <options> ;
  BY variables ;
  CDFPLOT <variables> < / options> ;
  CLASS variable-1 <(v-options)> < variable-2 <(v-options)> >
    < / KEYLEVEL= value1 | (value1 value2) > ;
  FREQ variable ;
  HISTOGRAM <variables> < / options> ;
  ID variables ;
  INSET keyword-list < / options> ;
  OUTPUT <OUT=SAS-data-set> <keyword1=names . . . keywordk=names> <percentile-options> ;
  PPPLOT <variables> < / options> ;
  PROBPLOT <variables> < / options> ;
  QQPLOT <variables> < / options> ;
  VAR variables ;
  WEIGHT variable ;
```

The PROC UNIVARIATE statement invokes the procedure. The VAR statement specifies the numeric variables to be analyzed, and it is required if the OUTPUT statement is used to save summary statistics in an output data set. If you do not use the VAR statement, all numeric variables in the data set are analyzed. The plot statements CDFPLOT, HISTOGRAM, PPPLOT, PROBPLOT, and QQPLOT create graphical displays, and the INSET statement enhances these displays by adding a table of summary statistics directly on the graph. You can specify one or more of each of the plot statements, the INSET statement, and the OUTPUT statement. If you use a VAR statement, the variables listed in a plot statement must be a subset of the variables listed in the VAR statement.

You can use a CLASS statement to specify one or two variables that group the data into classification levels. The analysis is carried out for each combination of levels. You can use the CLASS statement with plot statements to create comparative displays, in which each cell contains a plot for one combination of classification levels.

You can specify a BY statement to obtain separate analyses for each BY group. The FREQ statement specifies a variable whose values provide the frequency for each observation. The WEIGHT statement specifies a variable whose values are used to weight certain statistics. The ID statement specifies one or more variables to identify the extreme observations.

## PROC UNIVARIATE Statement

```
PROC UNIVARIATE <options> ;
```

The PROC UNIVARIATE statement is required to invoke the UNIVARIATE procedure. You can use the PROC UNIVARIATE statement by itself to request a variety of statistics for summarizing the data distribution of each analysis variable:

- sample moments
- basic measures of location and variability
• confidence intervals for the mean, standard deviation, and variance
• tests for location
• tests for normality
• trimmed and Winsorized means
• robust estimates of scale
• quantiles and related confidence intervals
• extreme observations and extreme values
• frequency counts for observations
• missing values

In addition, you can use options in the PROC UNIVARIATE statement to do the following:

• specify the input data set to be analyzed
• specify a graphics catalog for saving traditional graphics output
• specify rounding units for variable values
• specify the definition used to calculate percentiles
• specify the divisor used to calculate variances and standard deviations
• request that plots be produced on line printers and define special printing characters used for features
• suppress tables
• save statistics in an output data set

The following are the options that can be used with the PROC UNIVARIATE statement:

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 generated by the CIPCTLDF, CIPCTLNORMAL, LOCCOUNT, NORMAL, ROBUSTSCALE, TRIMMED=.25, and WINSORIZED=.25 options. PROC UNIVARIATE also uses any values that you specify for ALPHA=, MU0=, NEXTRVAL=, CIBASIC, CIPCTLDF, CIPCTLNORMAL, TRIMMED=, or WINSORIZED= to produce the output.

ALPHA=\alpha
specifies the level of significance \( \alpha \) for 100(1 – \( \alpha \))% confidence intervals. The value \( \alpha \) must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals.

Note that specialized ALPHA= options are available for a number of confidence interval options. For example, you can specify CIBASIC(ALPHA=0.10) to request a table of basic confidence limits at the 90% level. The default value of these options is the value of the ALPHA= option in the PROC statement.
ANNOTATE=SAS-data-set

specifies an input data set that contains annotate variables as described in SAS/GRAPH: Reference. You can use this data set to add features to your traditional graphics. PROC UNIVARIATE adds the features in this data set to every graph that is produced in the procedure. PROC UNIVARIATE does not use the ANNOTATE= data set unless you create a traditional graph with a plot statement. The option does not apply to ODS Graphics output. Use the ANNOTATE= option in the plot statement if you want to add a feature to a specific graph produced by that statement.

CIBASIC < (<TYPE=keyword > < ALPHA=α >)> 
requests confidence limits for the mean, standard deviation, and variance based on the assumption that the data are normally distributed. If you use the CIBASIC option, you must use the default value of VARDEF=, which is DF.

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

ALPHA=α
specifies the level of significance α for 100(1 − α)% confidence intervals. The value α must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

CIPCTLDF < (<TYPE=keyword > < ALPHA=α >)> 
CIQUANTDF < (<TYPE=keyword > < ALPHA=α >)> 
requests confidence limits for quantiles based on 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). This option does not apply if you use a WEIGHT statement.

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

ALPHA=α
specifies the level of significance α for 100(1 − α)% confidence intervals. The value α must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

CIPCTLNORMAL < (<TYPE=keyword > < ALPHA=α >)> 
CIQUANTNORMAL < (<TYPE=keyword > < ALPHA=α >)> 
requests confidence limits for quantiles based on the assumption that the data are normally distributed. The computational method is described in Section 4.4.1 of Hahn and Meeker (1991) and uses the noncentral t distribution as given by Odeh and Owen (1980). This option does not apply if you use a WEIGHT statement.

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

specifies the level of significance $\alpha$ for 100(1 - $\alpha$)% confidence intervals. The value $\alpha$ must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

**DATA=**SAS-data-set

specifies the input SAS data set to be analyzed. If the DATA= option is omitted, the procedure uses the most recently created SAS data set.

**EXCLNPWGT**

**EXCLNPWGTS**

excludes observations with nonpositive weight values (zero or negative) from the analysis. By default, PROC UNIVARIATE counts observations with negative or zero weights in the total number of observations. This option applies only when you use a WEIGHT statement.

**FREQ**

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

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

**GOUT=**graphics-catalog

specifies the SAS catalog that PROC UNIVARIATE uses to save traditional graphics output. If you omit the libref in the name of the graphics-catalog, PROC UNIVARIATE looks for the catalog in the temporary library called WORK and creates the catalog if it does not exist. The option does not apply to ODS Graphics output.

**IDOUT**

includes ID variables in the output data set created by an OUTPUT statement. The value of an ID variable in the output data set is its first value from the input data set or BY group. By default, ID variables are not included in OUTPUT statement data sets.

**LOCCOUNT**

requests a table that shows the number of observations greater than, not equal to, and less than the value of MU0=. PROC UNIVARIATE uses these values to construct the sign test and the signed rank test. This option does not apply if you use a WEIGHT statement.

**MODES|MODE**

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.

**MU0=**values

**LOCATION=**values

specifies the value of the mean or location parameter ($\mu_0$) in the null hypothesis for tests of location summarized in the table labeled “Tests for Location: Mu0=value.” 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, matching variables and location values by their order in the two lists. The default value is 0.
The following statement tests the hypothesis $\mu_0 = 0$ for the first variable and the hypothesis $\mu_0 = 0.5$ for the second variable.

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

**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. The default value is 5. You can specify NEXTROBS=0 to suppress the table of extreme observations.

**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. By default, $n = 0$ and no table is displayed.

**NOBYPLOT**
suppresses side-by-side line printer box plots that are created by default when you use the BY statement and either 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. You can use the NOPRINT option in the HISTOGRAM statement to suppress the creation of its tables. Use NOPRINT when you want to create an OUT= or OUTTABLE= output data set only.

**NORMAL**
**NORMALTEST**
requests tests for normality that include a series of goodness-of-fit tests based on the empirical distribution function. The table provides test statistics and $p$-values for the Shapiro-Wilk test (provided the sample size is less than or equal to 2000), the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér–von Mises test. This option does not apply if you use a WEIGHT statement.

**NOTABCONTENTS**
suppresses the table of contents entries for tables of summary statistics produced by the PROC UNIVARIATE statement.

**NOVARCONTENTS**
suppresses grouping entries associated with analysis variables in the table of contents. By default, the table of contents lists results associated with an analysis variable in a group with the variable name.

**OUTTABLE=SAS-data-set**
creates an output data set that contains univariate statistics arranged in tabular form, with one observation per analysis variable. See the section “OUTTABLE= Output Data Set” on page 433 for details.

**PCTLDEF=value**
**DEF=value**
specifies the definition that PROC UNIVARIATE uses to calculate quantiles. The default value is 5. Values can be 1, 2, 3, 4, or 5. You cannot use PCTLDEF= when you compute weighted quantiles. See the section “Calculating Percentiles” on page 382 for details on quantile definitions.
PROC UNIVARIATE Statement ♦ 279

PLOTS | PLOT< ( <plot-options> <SSPLOT(plot-options)> ) >
produces a panel of plots for each analysis variable. If ODS Graphics is enabled, the panel contains a horizontal histogram, a box plot, and a normal probability plot. Otherwise, the procedure produces a stem-and-leaf plot (or a horizontal bar chart), a box plot, and a normal probability plot by using line printer output. If you specify a BY statement, side-by-side box plots of the data from the BY groups are displayed following the univariate output for the last BY group.

You can specify the following plot options to produce titles and footnotes for the plots when ODS Graphics is enabled. Plot options that are specified within the SSPLOT suboption apply to the side-by-side box plots of BY-group data.

**ODSFOOTNOTE**=FOOTNOTE | FOOTNOTE1 | ‘string’
adds a footnote to ODS Graphics output. If you specify the FOOTNOTE (or FOOTNOTE1) keyword, the value of the SAS FOOTNOTE statement is used as the graph footnote. If you specify a quoted string, that is used as the footnote. The quoted string can contain either of the following escaped characters, which are replaced with the appropriate values from the analysis:

\n is replaced by the analysis variable name.
\l is replaced by the analysis variable label (or name if the analysis variable has no label).

**ODSFOOTNOTE2**=FOOTNOTE2 | ‘string’
adds a secondary footnote to ODS Graphics output. If you specify the FOOTNOTE2 keyword, the value of the SAS FOOTNOTE2 statement is used as the secondary graph footnote. If you specify a quoted string, that is used as the secondary footnote. The quoted string can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

\n is replaced by the analysis variable name.
\l is replaced by the analysis variable label (or name if the analysis variable has no label).

**ODSTITLE**=TITLE | TITLE1 | NONE | DEFAULT | LABELFMT | ‘string’
specifies a title for ODS Graphics output.

TITLE (or TITLE1) uses the value of SAS TITLE statement as the graph title.
NONE suppresses all titles from the graph.
DEFAULT uses the default ODS Graphics title (a descriptive title that consists of the plot type and the analysis variable name).
LABELFMT uses the default ODS Graphics title with the variable label instead of the variable name.

If you specify a quoted string, that is used as the graph title. The quoted string can contain the following escaped characters, which are replaced with the appropriate values from the analysis:
Chapter 4: The UNIVARIATE Procedure

\n is replaced by the analysis variable name.
\l is replaced by the analysis variable label (or name if the analysis variable has no label).

**ODSTITLE2=TITLE2 | ‘string’**
specifies a secondary title for ODS Graphics output. If you specify the TITLE2 keyword, the value of the SAS TITLE2 statement is used as the secondary graph title. If you specify a quoted string, that is used as the secondary title. The quoted string can contain the following escaped characters, which are replaced with the appropriate values from the analysis:

\n is replaced by the analysis variable name.
\l is replaced by the analysis variable label (or name if the analysis variable has no label).

**NOTE:** ODSTITLE=LABELFMT and the substitution of analysis variable names and labels are not supported for plot options specified within the SSMLOT suboption.

**PLOTSIZE=n**
specifies the approximate number of rows used in line-printer plots requested with the PLOTS option. 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 8, PROC UNIVARIATE uses eight rows to draw the plots.

**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 (\(\text{MAD} \)), and two statistics proposed by Rousseeuw and Croux (1993), \( Q_n \), and \( S_n \). See the section “Robust Estimates of Scale” on page 388 for details. This option does not apply if you use a WEIGHT statement.

**ROUND=units**
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. The ROUND= option reduces the number of unique variable values, thereby reducing memory requirements for the procedure. For example, to make the rounding unit 1 for the first analysis variable and 0.5 for the second analysis variable, submit the statement

```
proc univariate round=1 0.5;
   var Yieldstrength tenstren;
run;
```

When a variable value is midway between the two nearest rounded points, the value is rounded to the nearest even multiple of the roundoff value. For example, with a roundoff value of 1, the variable values of \(-2.5, -2.2, \) and \(-1.5\) are rounded to \(-2\); the values of \(-0.5, 0.2, \) and \(0.5\) are rounded to \(0\); and the values of \(0.6, 1.2, \) and \(1.4\) are rounded to \(1\).
SUMMARYCONTENTS='string'
specifies the table of contents entry used for grouping the summary statistics produced by the PROC UNIVARIATE statement. You can specify SUMMARYCONTENTS='' to suppress the grouping entry.

TRIMMED=values (< TYPE=keyword > < ALPHA=α >)
requests a table of trimmed means, where value specifies the number or the proportion of observations that PROC UNIVARIATE trims. If the value is the number n of trimmed observations, n must be between 0 and half the number of nonmissing observations. If value is a proportion p between 0 and 1/2, 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. To include confidence limits for the mean and the Student’s t test in the table, you must use the default value of VARDEF=, which is DF. For details concerning the computation of trimmed means, see the section “Trimmed Means” on page 388. The TRIMMED= option does not apply if you use a WEIGHT statement.

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

ALPHA=α
specifies the level of significance α for 100(1 − α)% confidence intervals. The value α must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals.

VARDEF=divisor
specifies the divisor to use in the calculation of variances and standard deviation. By default, VARDEF=DF. Table 4.1 shows the possible values for divisor and associated divisors.

<table>
<thead>
<tr>
<th>Value</th>
<th>Divisor</th>
<th>Formula for Divisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>degrees of freedom</td>
<td>n − 1</td>
</tr>
<tr>
<td>N</td>
<td>number of observations</td>
<td>n</td>
</tr>
<tr>
<td>WDF</td>
<td>sum of weights minus one</td>
<td>(Σ_i wi) − 1</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>sum of weights</td>
<td>Σ_i wi</td>
</tr>
</tbody>
</table>

The procedure computes the variance as \( \frac{CSS}{divisor} \) where CSS is the corrected sums of squares and equals \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \). When you weight the analysis variables, \( CSS = \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 \) where \( \bar{x}_w \) is the weighted mean.

The default value is DF. To compute the standard error of the mean, confidence limits, and Student’s t test, use the default value of VARDEF=.

When you use the WEIGHT statement and VARDEF=DF, the variance is an estimate of \( \sigma^2 \) where the variance of the \( i \)th observation is \( var(x_i) = \frac{\sigma^2}{w_i} \) and \( w_i \) is the weight for the \( i \)th observation. This yields an estimate of the variance of an observation with unit weight.

When you use the WEIGHT statement and VARDEF=WGT, the computed variance is asymptotically (for large n) an estimate of \( \frac{\sigma^2}{\bar{w}} \) where \( \bar{w} \) is the average weight. This yields an asymptotic estimate of the variance of an observation with average weight.
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 the value is the number \( n \) of Winsorized observations, \( n \) must be between 0 and half the number of nonmissing observations. If value is a proportion \( p \) between 0 and \( \frac{1}{2} \), 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. To include confidence limits for the mean and the Student \( t \) test in the table, you must use the default value of \( \text{VARDEF=} \), which is DF. For details concerning the computation of Winsorized means, see the section “Winsorized Means” on page 387. The WINSORIZED= option does not apply if you use a WEIGHT statement.

\[ \text{TYPE=} \text{keyword} \]

- \text{keyword} specifies the type of confidence limit for the mean, where \text{keyword} is LOWER, UPPER, or TWOSIDED. The default is TWOSIDED.

\[ \text{ALPHA=} \alpha \]

- \( \alpha \) specifies the level of significance \( \alpha \) for 100(1 - \( \alpha \))% confidence intervals. The value \( \alpha \) must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals.

### BY Statement

**BY** variables ;

You can specify a BY statement with PROC UNIVARIATE to obtain separate analyses for each BY group. The BY statement specifies the variables 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 be indexed appropriately.

**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.

### CDFPLOT Statement

**CDFPLOT** < variables > < / options > ;
The CDFPLOT statement plots the observed cumulative distribution function (cdf) of a variable, defined as

\[
F_N(x) = \frac{\text{percent of nonmissing values } \leq x}{N} \times 100\%
\]

where \( N \) is the number of nonmissing observations. The cdf is an increasing step function that has a vertical jump of \( \frac{1}{N} \) at each value of \( x \) equal to an observed value. The cdf is also referred to as the empirical cumulative distribution function (ECDF).

You can use any number of CDFPLOT statements in the UNIVARIATE procedure. The components of the CDFPLOT statement are as follows.

**variables**

specify variables for which to create cdf plots. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify a list of variables, then by default the procedure creates a cdf plot for each variable listed in the VAR statement, or for each numeric variable in the DATA= data set if you do not specify a VAR statement.

For example, suppose a data set named Steel contains exactly three numeric variables: Length, Width, and Height. The following statements create a cdf plot for each of the three variables:

```plaintext
proc univariate data=Steel;
    cdfplot;
run;
```

The following statements create a cdf plot for Length and a cdf plot for Width:

```plaintext
proc univariate data=Steel;
    var Length Width;
    cdfplot;
run;
```

The following statements create a cdf plot for Width:

```plaintext
proc univariate data=Steel;
    var Length Width;
    cdfplot Width;
run;
```

**options**

specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all options after the slash (/) in the CDFPLOT statement. You can specify only one option that names a distribution in each CDFPLOT statement, but you can specify any number of other options. The distributions available are listed in Table 4.2. By default, the procedure produces a plot for the normal distribution.

Table 4.2 through Table 4.4 list the CDFPLOT options by function. For complete descriptions, see the sections “Dictionary of Options” on page 288 and “Dictionary of Common Options” on page 369. Options can be any of the following:
• primary options
• secondary options
• general options

Distribution Options

Table 4.2 lists primary options for requesting a theoretical distribution.

Table 4.2  Primary Options for Theoretical Distribution

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA(beta-options)</td>
<td>plots two-parameter beta distribution function, parameters $\theta$ and $\sigma$ assumed known</td>
</tr>
<tr>
<td>EXPONENTIAL(exponential-options)</td>
<td>plots one-parameter exponential distribution function, parameter $\theta$ assumed known</td>
</tr>
<tr>
<td>GAMMA(gamma-options)</td>
<td>plots two-parameter gamma distribution function, parameter $\theta$ assumed known</td>
</tr>
<tr>
<td>GUMBEL(Gumbel-options)</td>
<td>plots Gumbel distribution with location parameter $\mu$ and scale parameter $\sigma$</td>
</tr>
<tr>
<td>IGAUSS(iGauss-options)</td>
<td>plots inverse Gaussian distribution with mean $\mu$ and shape parameter $\lambda$</td>
</tr>
<tr>
<td>LOGNORMAL(lognormal-options)</td>
<td>plots two-parameter lognormal distribution function, parameter $\theta$ assumed known</td>
</tr>
<tr>
<td>NORMAL(normal-options)</td>
<td>plots normal distribution function</td>
</tr>
<tr>
<td>PARETO(Pareto-options)</td>
<td>plots generalized Pareto distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameter $\alpha$</td>
</tr>
<tr>
<td>POWER(power-options)</td>
<td>plots power function distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameter $\alpha$</td>
</tr>
<tr>
<td>RAYLEIGH(Rayleigh-options)</td>
<td>plots Rayleigh distribution with threshold parameter $\theta$ and scale parameter $\sigma$</td>
</tr>
<tr>
<td>WEIBULL(Weibull-options)</td>
<td>plots two-parameter Weibull distribution function, parameter $\theta$ assumed known</td>
</tr>
</tbody>
</table>

Table 4.3 lists secondary options that specify distribution parameters and control the display of a theoretical distribution function. Specify these options in parentheses after the distribution keyword. For example, you can request a normal probability plot with a distribution reference line by specifying the NORMAL option as follows:
The COLOR= option specifies the color for the curve, and the normal-options MU= and SIGMA= specify the parameters $\mu = 10$ and $\sigma = 0.5$ for the distribution function. If you do not specify these parameters, maximum likelihood estimates are computed.

### Table 4.3 Secondary Distribution Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Options Used with All Distributions</strong></td>
<td></td>
</tr>
<tr>
<td>COLOR=</td>
<td>specifies color of theoretical distribution function</td>
</tr>
<tr>
<td>L=</td>
<td>specifies line type of theoretical distribution function</td>
</tr>
<tr>
<td>W=</td>
<td>specifies width of theoretical distribution function</td>
</tr>
<tr>
<td><strong>Beta-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies first shape parameter $\alpha$ for beta distribution function</td>
</tr>
<tr>
<td>BETA=</td>
<td>specifies second shape parameter $\beta$ for beta distribution function</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for beta distribution function</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies lower threshold parameter $\theta$ for beta distribution function</td>
</tr>
<tr>
<td><strong>Exponential-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for exponential distribution function</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for exponential distribution function</td>
</tr>
<tr>
<td><strong>Gamma-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$ for gamma distribution function</td>
</tr>
<tr>
<td>ALPHADELTA=</td>
<td>specifies change in successive estimates of $\alpha$ at which the Newton-Raphson approximation of $\hat{\alpha}$ terminates</td>
</tr>
<tr>
<td>ALPHAINITIAL=</td>
<td>specifies initial value for $\alpha$ in the Newton-Raphson approximation of $\alpha$</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{\alpha}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for gamma distribution function</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for gamma distribution function</td>
</tr>
<tr>
<td><strong>Gumbel-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies location parameter $\mu$ for Gumbel distribution function</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for Gumbel distribution function</td>
</tr>
<tr>
<td><strong>IGauss-Options</strong></td>
<td></td>
</tr>
<tr>
<td>LAMBDA=</td>
<td>specifies shape parameter $\lambda$ for inverse Gaussian distribution function</td>
</tr>
<tr>
<td>MU=</td>
<td>specifies mean $\mu$ for inverse Gaussian distribution function</td>
</tr>
<tr>
<td><strong>Lognormal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies shape parameter $\sigma$ for lognormal distribution function</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for lognormal distribution function</td>
</tr>
<tr>
<td>ZETA=</td>
<td>specifies scale parameter $\zeta$ for lognormal distribution function</td>
</tr>
<tr>
<td><strong>Normal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies mean $\mu$ for normal distribution function</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies standard deviation $\sigma$ for normal distribution function</td>
</tr>
<tr>
<td><strong>Pareto-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$ for generalized Pareto distribution function</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for generalized Pareto distribution function</td>
</tr>
</tbody>
</table>
### Table 4.3 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter ( \theta ) for generalized Pareto distribution function</td>
</tr>
<tr>
<td><strong>Power-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter ( \alpha ) for power function distribution</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter ( \sigma ) for power function distribution</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter ( \theta ) for power function distribution</td>
</tr>
<tr>
<td><strong>Rayleigh-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter ( \sigma ) for Rayleigh distribution function</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter ( \theta ) for Rayleigh distribution function</td>
</tr>
<tr>
<td><strong>Secondary Weibull-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies shape parameter ( c ) for Weibull distribution function</td>
</tr>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approxima-</td>
</tr>
<tr>
<td></td>
<td>tion of ( \hat{c} )</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter ( \sigma ) for Weibull distribution function</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter ( \theta ) for Weibull distribution function</td>
</tr>
</tbody>
</table>

### General Options

Table 4.4 summarizes general options for enhancing cdf plots.

### Table 4.4 General Graphics Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Graphics Options</strong></td>
<td></td>
</tr>
<tr>
<td>HREF=</td>
<td>specifies reference lines perpendicular to the horizontal axis</td>
</tr>
<tr>
<td>HREFLABELS=</td>
<td>specifies labels for HREF= lines</td>
</tr>
<tr>
<td>HREFLABPOS=</td>
<td>specifies position for HREF= line labels</td>
</tr>
<tr>
<td>NOEcdf</td>
<td>suppresses plot of empirical (observed) distribution function</td>
</tr>
<tr>
<td>NOHLABEL</td>
<td>suppresses label for horizontal axis</td>
</tr>
<tr>
<td>NOvLABEL</td>
<td>suppresses label for vertical axis</td>
</tr>
<tr>
<td>NOvICK</td>
<td>suppresses tick marks and tick mark labels for vertical axis</td>
</tr>
<tr>
<td>STATREF=</td>
<td>specifies reference lines at values of summary statistics</td>
</tr>
<tr>
<td>STATREFLABELS=</td>
<td>specifies labels for STATREF= lines</td>
</tr>
<tr>
<td>STATREFSUBCHAR=</td>
<td>specifies substitution character for displaying statistic values in STA-</td>
</tr>
<tr>
<td></td>
<td>TREFLABELS= lines</td>
</tr>
<tr>
<td>VAXISLABEL=</td>
<td>specifies label for vertical axis</td>
</tr>
<tr>
<td>VREF=</td>
<td>specifies reference lines perpendicular to the vertical axis</td>
</tr>
<tr>
<td>VREFLABELS=</td>
<td>specifies labels for VREF= lines</td>
</tr>
<tr>
<td>VREFLABPOS=</td>
<td>specifies position for VREF= line labels</td>
</tr>
<tr>
<td>VSSCALE=</td>
<td>specifies scale for vertical axis</td>
</tr>
<tr>
<td><strong>Options for Traditional Graphics Output</strong></td>
<td></td>
</tr>
<tr>
<td>ANNOTATE=</td>
<td>specifies annotate data set</td>
</tr>
<tr>
<td>CAXIS=</td>
<td>specifies color for axis</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------------------------------------------</td>
</tr>
<tr>
<td>CFRAME=</td>
<td>specifies color for frame</td>
</tr>
<tr>
<td>CHREF=</td>
<td>specifies colors for HREF= lines</td>
</tr>
<tr>
<td>CSTATREF=</td>
<td>specifies colors for STATREF= lines</td>
</tr>
<tr>
<td>CTEXT=</td>
<td>specifies color for text</td>
</tr>
<tr>
<td>CVREF=</td>
<td>specifies colors for VREF= lines</td>
</tr>
<tr>
<td>DESCRIPTION=</td>
<td>specifies description for graphics catalog member</td>
</tr>
<tr>
<td>FONT=</td>
<td>specifies text font</td>
</tr>
<tr>
<td>HAXIS=</td>
<td>specifies AXIS statement for horizontal axis</td>
</tr>
<tr>
<td>HEIGHT=</td>
<td>specifies height of text used outside framed areas</td>
</tr>
<tr>
<td>HMINOR=</td>
<td>specifies number of horizontal axis minor tick marks</td>
</tr>
<tr>
<td>INFONT=</td>
<td>specifies software font for text inside framed areas</td>
</tr>
<tr>
<td>INHEIGHT=</td>
<td>specifies height of text inside framed areas</td>
</tr>
<tr>
<td>LHREF=</td>
<td>specifies line types for HREF= lines</td>
</tr>
<tr>
<td>LSTATREF=</td>
<td>specifies line types for STATREF= lines</td>
</tr>
<tr>
<td>LVREF=</td>
<td>specifies line types for VREF= lines</td>
</tr>
<tr>
<td>NAME=</td>
<td>specifies name for plot in graphics catalog</td>
</tr>
<tr>
<td>NOFRAME</td>
<td>suppresses frame around plotting area</td>
</tr>
<tr>
<td>TURNVLABELS</td>
<td>turns and vertically strings out characters in labels for vertical axis</td>
</tr>
<tr>
<td>VAXIS=</td>
<td>specifies AXIS statement for vertical axis</td>
</tr>
<tr>
<td>VMINOR=</td>
<td>specifies number of vertical axis minor tick marks</td>
</tr>
<tr>
<td>WAXIS=</td>
<td>specifies line thickness for axes and frame</td>
</tr>
</tbody>
</table>

**Options for ODS Graphics Output**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODSFOOTNOTE=</td>
<td>specifies footnote displayed on plot</td>
</tr>
<tr>
<td>ODSFOOTNOTE2=</td>
<td>specifies secondary footnote displayed on plot</td>
</tr>
<tr>
<td>ODSTITLE=</td>
<td>specifies title displayed on plot</td>
</tr>
<tr>
<td>ODSTITLE2=</td>
<td>specifies secondary title displayed on plot</td>
</tr>
<tr>
<td>OVERLAY</td>
<td>overlays plots for different class levels</td>
</tr>
</tbody>
</table>

**Options for Comparative Plots**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNOKEY</td>
<td>applies annotation requested in ANNOTATE= data set to key cell only</td>
</tr>
<tr>
<td>CFrameside=</td>
<td>specifies color for filling row label frames</td>
</tr>
<tr>
<td>CFRAMETOP=</td>
<td>specifies color for filling column label frames</td>
</tr>
<tr>
<td>CPROP=</td>
<td>specifies color for proportion of frequency bar</td>
</tr>
<tr>
<td>CTEXTSIDE=</td>
<td>specifies color for row labels</td>
</tr>
<tr>
<td>CTEXTTOP=</td>
<td>specifies color for column labels</td>
</tr>
<tr>
<td>INTERTILE=</td>
<td>specifies distance between tiles in comparative plot</td>
</tr>
<tr>
<td>NCOLS=</td>
<td>specifies number of columns in comparative plot</td>
</tr>
<tr>
<td>NROWS=</td>
<td>specifies number of rows in comparative plot</td>
</tr>
</tbody>
</table>

**Miscellaneous Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTENTS=</td>
<td>specifies table of contents entry for cdf plot grouping</td>
</tr>
</tbody>
</table>
Dictionary of Options

The following entries provide detailed descriptions of the options specific to the CDFPLOT statement. See the section “Dictionary of Common Options” on page 369 for detailed descriptions of options common to all plot statements.

**ALPHA=value**

specifies the shape parameter \( \alpha \) for distribution functions requested with the BETA, GAMMA, PARETO, and POWER options. Enclose the ALPHA= option in parentheses after the distribution keyword. If you do not specify a value for \( \alpha \), the procedure calculates a maximum likelihood estimate. For examples, see the entries for the BETA and GAMMA options.

**BETA<(beta-options)>**

displays a fitted beta distribution function on the cdf plot. The equation of the fitted cdf is

\[
F(x) = \begin{cases} 
0 & \text{for } x \leq \theta \\
\frac{I_{\frac{x-\theta}{\sigma}}(\alpha, \beta)}{I_{\frac{\theta}{\sigma}}(\alpha, \beta)} & \text{for } \theta < x < \theta + \sigma \\
1 & \text{for } x \geq \sigma + \theta 
\end{cases}
\]

where \( I_y(\alpha, \beta) \) is the incomplete beta function and

\[
\begin{align*}
\theta &= \text{lower threshold parameter (lower endpoint)} \\
\sigma &= \text{scale parameter } (\sigma > 0) \\
\alpha &= \text{shape parameter } (\alpha > 0) \\
\beta &= \text{shape parameter } (\beta > 0)
\end{align*}
\]

The beta distribution is bounded below by the parameter \( \theta \) and above by the value \( \theta + \sigma \). You can specify \( \theta \) and \( \sigma \) by using the THETA= and SIGMA= beta-options, as illustrated in the following statements, which fit a beta distribution bounded between 50 and 75. The default values for \( \theta \) and \( \sigma \) are 0 and 1, respectively.

```plaintext
proc univariate;
  cdfplot / beta(theta=50 sigma=25);
run;
```

The beta distribution has two shape parameters: \( \alpha \) and \( \beta \). If these parameters are known, you can specify their values with the ALPHA= and BETA= beta-options. If you do not specify values for \( \alpha \) and \( \beta \), the procedure calculates maximum likelihood estimates.

The BETA option can appear only once in a CDFPLOT statement. Table 4.3 lists options you can specify with the BETA distribution option.

**BETA=value**

**B=value**

specifies the second shape parameter \( \beta \) for beta distribution functions requested by the BETA option. Enclose the BETA= option in parentheses after the BETA keyword. If you do not specify a value for \( \beta \), the procedure calculates a maximum likelihood estimate. For examples, see the preceding entry for the BETA option.
C=value

specifies the shape parameter $c$ for Weibull distribution functions requested with the WEIBULL option. Enclose the C= option in parentheses after the WEIBULL keyword. If you do not specify a value for $c$, the procedure calculates a maximum likelihood estimate. You can specify the SHAPE= option as an alias for the C= option.

**EXPONENTIAL**: (exponential-options)

**EXP**: (exponential-options)

displays a fitted exponential distribution function on the cdf plot. The equation of the fitted cdf is

\[
F(x) = \begin{cases} 
0 & \text{for } x \leq \theta \\
1 - \exp \left( -\frac{x-\theta}{\sigma} \right) & \text{for } x > \theta
\end{cases}
\]

where

\[
\theta = \text{threshold parameter} \\
\sigma = \text{scale parameter (} \sigma > 0\text{)}
\]

The parameter $\theta$ must be less than or equal to the minimum data value. You can specify $\theta$ with the THETA= exponential-option. The default value for $\theta$ is 0. You can specify $\sigma$ with the SIGMA= exponential-option. By default, a maximum likelihood estimate is computed for $\sigma$. For example, the following statements fit an exponential distribution with $\theta = 10$ and a maximum likelihood estimate for $\sigma$:

```plaintext
proc univariate;
  cdfplot / exponential(theta=10 l=2 color=green);
run;
```

The exponential curve is green and has a line type of 2.

The EXPONENTIAL option can appear only once in a CDFPLOT statement. Table 4.3 lists the options you can specify with the EXPONENTIAL option.

**GAMMA**: (gamma-options)

displays a fitted gamma distribution function on the cdf plot. The equation of the fitted cdf is

\[
F(x) = \begin{cases} 
0 & \text{for } x \leq \theta \\
\frac{1}{\Gamma(\alpha)\sigma} \int_0^x \left( \frac{t-\theta}{\sigma} \right)^{\alpha-1} exp \left(-\frac{t-\theta}{\sigma}\right) dt & \text{for } x > \theta
\end{cases}
\]

where

\[
\theta = \text{threshold parameter} \\
\sigma = \text{scale parameter (} \sigma > 0\text{)} \\
\alpha = \text{shape parameter (} \alpha > 0\text{)}
\]

The parameter $\theta$ for the gamma distribution must be less than the minimum data value. You can specify $\theta$ with the THETA= gamma-option. The default value for $\theta$ is 0. In addition, the gamma distribution has a shape parameter $\alpha$ and a scale parameter $\sigma$. You can specify these parameters with the ALPHA= and SIGMA= gamma-options. By default, maximum likelihood estimates are computed for $\alpha$ and $\sigma$. For example, the following statements fit a gamma distribution function with $\theta = 4$ and maximum likelihood estimates for $\alpha$ and $\sigma$:

```plaintext
proc univariate;
  cdfplot / gamma(theta=4 alpha l=2 color=green);
run;
```
proc univariate;
  cdfplot / gamma(theta=4);
run;

Note that the maximum likelihood estimate of $\alpha$ is calculated iteratively using the Newton-Raphson approximation. The gamma-options ALPHADELTA=, ALPHAINITIAL=, and MAXITER= control the approximation.

The GAMMA option can appear only once in a CDFPLOT statement. Table 4.3 lists the options you can specify with the GAMMA option.

**GUMBEL<(Gumbel-options)>**

displays a fitted Gumbel distribution (also known as Type 1 extreme value distribution) function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \exp \left( -e^{-(x-\mu)/\sigma} \right)$$

where

\[ \mu = \text{location parameter} \]
\[ \sigma = \text{scale parameter} \quad (\sigma > 0) \]

You can specify known values for $\mu$ and $\sigma$ with the MU= and SIGMA= Gumbel-options. By default, maximum likelihood estimates are computed for $\mu$ and $\sigma$.

The GUMBEL option can appear only once in a CDFPLOT statement. Table 4.3 lists secondary options you can specify with the GUMBEL option.

**IGAUSS<(iGauss-options)>**

displays a fitted inverse Gaussian distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \Phi \left\{ \sqrt{\frac{\lambda}{x}} \left( \frac{x}{\mu} - 1 \right) \right\} + e^{2\lambda/\mu} \Phi \left\{ -\sqrt{\frac{\lambda}{x}} \left( \frac{x}{\mu} + 1 \right) \right\}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

\[ \mu = \text{mean parameter} \quad (\mu > 0) \]
\[ \lambda = \text{shape parameter} \quad (\lambda > 0) \]

You can specify known values for $\mu$ and $\lambda$ with the MU= and LAMBDA= iGauss-options. By default, maximum likelihood estimates are computed for $\mu$ and $\lambda$.

The IGAUSS option can appear only once in a CDFPLOT statement. Table 4.3 lists secondary options you can specify with the IGAUSS option.

**LAMBDA=value**

specifies the shape parameter $\lambda$ for distribution functions requested with the IGAUSS option. Enclose the LAMBDA= option in parentheses after the IGAUSS distribution keyword. If you do not specify a value for $\lambda$, the procedure calculates a maximum likelihood estimate.
LOGNORMAL< (lognormal-options) >
displays a fitted lognormal distribution function on the cdf plot. The equation of the fitted cdf is
\[
F(x) = \begin{cases} 
0 & \text{for } x \leq \theta \\
\Phi \left( \frac{\log(x-\theta) - \zeta}{\sigma} \right) & \text{for } x > \theta 
\end{cases}
\]
where \(\Phi(\cdot)\) is the standard normal cumulative distribution function and
\[
\begin{align*}
\theta &= \text{threshold parameter} \\
\zeta &= \text{scale parameter} \\
\sigma &= \text{shape parameter } (\sigma > 0)
\end{align*}
\]
The parameter \(\theta\) for the lognormal distribution must be less than the minimum data value. You can specify \(\theta\) with the THETA= lognormal-option. The default value for \(\theta\) is 0. In addition, the lognormal distribution has a shape parameter \(\sigma\) and a scale parameter \(\zeta\). You can specify these parameters with the SIGMA= and ZETA= lognormal-options. By default, maximum likelihood estimates are computed for \(\sigma\) and \(\zeta\). For example, the following statements fit a lognormal distribution function with \(\theta = 10\) and maximum likelihood estimates for \(\sigma\) and \(\zeta\):

```proc univariate;
   cdfplot / lognormal(theta = 10);
run;
```

The LOGNORMAL option can appear only once in a CDFPLOT statement. Table 4.3 lists options that you can specify with the LOGNORMAL option.

**MU=value**
specifies the parameter \(\mu\) for theoretical cumulative distribution functions requested with the GUMBEL, IGAUSS, and NORMAL option. Enclose the MU= option in parentheses after the distribution keyword. For the inverse Gaussian and normal distributions, the default value is the sample mean. If you do not specify a value for \(\mu\) for the Gumbel distribution, the procedure calculates a maximum likelihood estimate. For an example, see the entry for the NORMAL option.

**NOECDF**
suppresses the observed distribution function (the empirical cumulative distribution function) of the variable, which is drawn by default. This option enables you to create theoretical cdf plots without displaying the data distribution. The NOECDF option can be used only with a theoretical distribution (such as the NORMAL option).

**NORMAL< (normal-options) >**
displays a fitted normal distribution function on the cdf plot. The equation of the fitted cdf is
\[
F(x) = \Phi \left( \frac{x-\mu}{\sigma} \right) \quad \text{for } -\infty < x < \infty
\]
where \(\Phi(\cdot)\) is the standard normal cumulative distribution function and
\[
\begin{align*}
\mu &= \text{mean} \\
\sigma &= \text{standard deviation } (\sigma > 0)
\end{align*}
\]
You can specify known values for \(\mu\) and \(\sigma\) with the MU= and SIGMA= normal-options, as shown in the following statements:
Chapter 4: The UNIVARIATE Procedure

```plaintext
proc univariate;
   cdfplot / normal(mu=14 sigma=.05);
run;
```

By default, the sample mean and sample standard deviation are calculated for \( \mu \) and \( \sigma \). The NORMAL option can appear only once in a CDFPLOT statement. Table 4.3 lists options that you can specify with the NORMAL option.

**PARETO**< (Pareto-options)>
displays a fitted generalized Pareto distribution function on the cdf plot. The equation of the fitted cdf is

\[
F(x) = 1 - \left( 1 - \frac{x - \theta}{\sigma} \right)^{\frac{1}{\alpha}}
\]

where

- \( \theta \) = threshold parameter
- \( \sigma \) = scale parameter (\( \sigma > 0 \))
- \( \alpha \) = shape parameter

The parameter \( \theta \) for the generalized Pareto distribution must be less than the minimum data value. You can specify \( \theta \) with the THETA= Pareto-option. The default value for \( \theta \) is 0. In addition, the generalized Pareto distribution has a shape parameter \( \alpha \) and a scale parameter \( \sigma \). You can specify these parameters with the ALPHA= and SIGMA= Pareto-options. By default, maximum likelihood estimates are computed for \( \alpha \) and \( \sigma \).

The PARETO option can appear only once in a CDFPLOT statement. Table 4.3 lists options that you can specify with the PARETO option.

**POWER**< (power-options)>
displays a fitted power function distribution on the cdf plot. The equation of the fitted cdf is

\[
F(x) = \begin{cases} 
0 & \text{for } x \leq \theta \\
\left( \frac{x - \theta}{\sigma} \right)^{\alpha} & \text{for } \theta < x < \theta + \sigma \\
1 & \text{for } x \geq \theta + \sigma 
\end{cases}
\]

where

- \( \theta \) = lower threshold parameter (lower endpoint)
- \( \sigma \) = scale parameter (\( \sigma > 0 \))
- \( \alpha \) = shape parameter (\( \alpha > 0 \))

The power function distribution is bounded below by the parameter \( \theta \) and above by the value \( \theta + \sigma \). You can specify \( \theta \) and \( \sigma \) by using the THETA= and SIGMA= power-options. The default values for \( \theta \) and \( \sigma \) are 0 and 1, respectively.

You can specify a value for the shape parameter, \( \alpha \), with the ALPHA= power-option. If you do not specify a value for \( \alpha \), the procedure calculates a maximum likelihood estimate.

The power function distribution is a special case of the beta distribution with its second shape parameter, \( \beta = 1 \).
The POWER option can appear only once in a CDFPLOT statement. Table 4.3 lists options that you can specify with the POWER option.

**RAYLEIGH** \( (Rayleigh\text{-}options) \)
displays a fitted Rayleigh distribution function on the cdf plot. The equation of the fitted cdf is

\[
F(x) = 1 - e^{-(x-\theta)^2/(2\sigma^2)}
\]

where

\[
\theta = \text{threshold parameter} \\
\sigma = \text{scale parameter (}\sigma > 0) 
\]

The parameter \( \theta \) for the Rayleigh distribution must be less than the minimum data value. You can specify \( \theta \) with the THETA= Rayleigh-option. The default value for \( \theta \) is 0. You can specify \( \sigma \) with the SIGMA= Rayleigh-option. By default, a maximum likelihood estimate is computed for \( \sigma \).

The RAYLEIGH option can appear only once in a CDFPLOT statement. Table 4.3 lists options that you can specify with the RAYLEIGH option.

**SIGMA=** \( \text{value} \mid \text{EST} \)
specifies the parameter \( \sigma \) for distribution functions requested by the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, and WEIBULL options. Enclose the SIGMA= option in parentheses after the distribution keyword. The following table summarizes the use of the SIGMA= option:

<table>
<thead>
<tr>
<th>Distribution Option</th>
<th>SIGMA= Specifies</th>
<th>Default Value</th>
<th>Alias</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA</td>
<td>scale parameter ( \sigma )</td>
<td>1</td>
<td>SCALE=</td>
</tr>
<tr>
<td>EXPONENTIAL</td>
<td>scale parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td>SCALE=</td>
</tr>
<tr>
<td>GAMMA</td>
<td>scale parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td>SCALE=</td>
</tr>
<tr>
<td>GUMBEL</td>
<td>scale parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
<tr>
<td>LOGNORMAL</td>
<td>shape parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td>SHAPE=</td>
</tr>
<tr>
<td>NORMAL</td>
<td>scale parameter ( \sigma )</td>
<td>standard deviation</td>
<td></td>
</tr>
<tr>
<td>PARETO</td>
<td>scale parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
<tr>
<td>POWER</td>
<td>scale parameter ( \sigma )</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>RAYLEIGH</td>
<td>scale parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
<tr>
<td>WEIBULL</td>
<td>scale parameter ( \sigma )</td>
<td>maximum likelihood estimate</td>
<td>SCALE=</td>
</tr>
</tbody>
</table>

**THETA=** \( \text{value} \mid \text{EST} \)
**THRESHOLD=** \( \text{value} \mid \text{EST} \)
specifies the lower threshold parameter \( \theta \) for theoretical cumulative distribution functions requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, and WEIBULL options. Enclose the THETA= option in parentheses after the distribution keyword. The default value is 0.

**VSCALE=** \( \text{PERCENT} \mid \text{PROPORTION} \)
specifies the scale of the vertical axis. The value PERCENT scales the data in units of percent of observations per data unit. The value PROPORTION scales the data in units of proportion of observations per data unit. The default is PERCENT.
WEIBULL< (Weibull-options) >
displays a fitted Weibull distribution function on the cdf plot. The equation of the fitted cdf is
\[ F(x) = \begin{cases} 0 & \text{for } x \leq \theta \\ 1 - \exp \left( - \left( \frac{x - \theta}{\sigma} \right)^c \right) & \text{for } x > \theta \end{cases} \]
where
\[ \theta = \text{threshold parameter} \]
\[ \sigma = \text{scale parameter } (\sigma > 0) \]
\[ c = \text{shape parameter } (c > 0) \]
The parameter \( \theta \) must be less than the minimum data value. You can specify \( \theta \) with the THETA= Weibull-option. The default value for \( \theta \) is 0. In addition, the Weibull distribution has a shape parameter \( c \) and a scale parameter \( \sigma \). You can specify these parameters with the SIGMA= and C= Weibull-options. By default, maximum likelihood estimates are computed for \( c \) and \( \sigma \). For example, the following statements fit a Weibull distribution function with \( \theta = 15 \) and maximum likelihood estimates for \( \sigma \) and \( c \):

```plaintext
proc univariate;
    cdfplot / weibull(theta=15);
run;
```
The WEIBULL option can appear only once in a CDFPLOT statement. Table 4.3 lists options that you can specify with the WEIBULL option.

ZETA=value
specifies a value for the scale parameter \( \xi \) for a lognormal distribution function requested with the LOGNORMAL option. Enclose the ZETA= option in parentheses after the LOGNORMAL keyword. If you do not specify a value for \( \xi \), a maximum likelihood estimate is computed. You can specify the SCALE= option as an alias for the ZETA= option.

---

**CLASS Statement**

CLASS variable-1 < (v-options) > < variable-2 < (v-options) >> < / KEYLEVEL= value1 | (value1 value2) > ;

The CLASS statement specifies one or two variables used 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 floating point 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 specify the following v-options enclosed in parentheses after the CLASS variable:
**MISSING**
specifies that missing values for the CLASS variable are to be treated as valid classification levels. Special missing values that represent numeric values (‘.A’ through ‘.Z’ and ‘._’) are each considered as a separate value. If you omit MISSING, PROC UNIVARIATE excludes the observations with a missing CLASS variable value from the analysis. Enclose this option in parentheses after the CLASS variable.

**ORDER=DATA | FORMATTED | FREQ | INTERNAL**
specifies the display order for the CLASS variable values. The default value is INTERNAL. You can specify the following values with the ORDER=option:

- **DATA** orders values according to their order in the input data set. When you use a plot 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 might depend on your operating environment. When you use a plot 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, suppose 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.

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. For numeric variables without an explicit format, the levels are ordered by their internal values.

- **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.

When you use a plot 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 may depend on your operating environment.

When you use a plot 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 is used to label the columns of the comparative plots (left to right). For example, suppose 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.

You can specify the following option after the slash (/) in the CLASS statement.
KEYLEVEL=value | ( value1 value2 )
specifies the key cells in comparative plots. For each plot, PROC UNIVARIATE first determines the
horizontal axis scaling for the key cell, and then extends the axis using the established tick interval to
accommodate the data ranges for the remaining cells, if necessary. 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 plot statement, KEYLEVEL=value identifies the
key cell as the level for which the CLASS variable is equal to value. By default, PROC UNIVARIATE
sorts the levels in the order determined by the ORDER= option, and 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).

If you specify two CLASS variables, use KEYLEVEL=(value1 value2) to identify the key cell as the
level for which CLASS variable n is equal to value n. By default, PROC UNIVARIATE sorts the levels
of the first CLASS variable in the order that is determined by its ORDER= option. Then, within each
of these levels, it sorts the levels of the second CLASS variable in the order that is determined by its
ORDER= option. 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 the first CLASS variable from top to bottom
and in the order of the second CLASS variable 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.

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

The KEYLEVEL= option has no effect unless you specify a plot statement.

NOKEYMOVE
specifies that the location of the key cell in a comparative plot be unchanged by the CLASS statement
KEYLEVEL= option. By default, the key cell is positioned as the first cell in a comparative plot.

The NOKEYMOVE option has no effect unless you specify a plot statement.

FREQ Statement

FREQ variable ;

The FREQ statement 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 the variable is not an integer, the SAS System truncates it. If the variable is less
than 1 or is missing, the procedure excludes that observation from the analysis. See Example 4.6.

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

HISTOGRAM Statement

HISTOGRAM < variables > < / options > ;
The HISTOGRAM statement creates histograms and optionally superimposes estimated parametric and nonparametric probability density curves. You cannot use the WEIGHT statement with the HISTOGRAM statement. You can use any number of HISTOGRAM statements after a PROC UNIVARIATE statement. The components of the HISTOGRAM statement are follows.

variables
are the variables for which histograms are to be created. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify variables in a VAR statement or in the HISTOGRAM statement, then by default, a histogram is created for each numeric variable in the DATA= data set. If you use a VAR statement and do not specify any variables in the HISTOGRAM statement, then by default, a histogram is created for each variable listed in the VAR statement.

For example, suppose a data set named Steel contains exactly two numeric variables named Length and Width. The following statements create two histograms, one for Length and one for Width:

```
proc univariate data=Steel;
    histogram;
run;
```

Likewise, the following statements create histograms for Length and Width:

```
proc univariate data=Steel;
    var Length Width;
    histogram;
run;
```

The following statements create a histogram for Length only:

```
proc univariate data=Steel;
    var Length Width;
    histogram Length;
run;
```

options
add features to the histogram. Specify all options after the slash (/) in the HISTOGRAM statement. Options can be one of the following:

- primary options for fitted parametric distributions and kernel density estimates
- secondary options for fitted parametric distributions and kernel density estimates
- general options for graphics and output data sets

For example, in the following statements, the NORMAL option displays a fitted normal curve on the histogram, the MIDPOINTS= option specifies midpoints for the histogram, and the CTEXT= option specifies the color of the text:

```
proc univariate data=Steel;
    histogram Length;
    normal;
run;
```
Table 4.5 through Table 4.8 list the HISTOGRAM options by function. For complete descriptions, see the sections “Dictionary of Options” on page 304 and “Dictionary of Common Options” on page 369.

### Parametric Density Estimation Options

Table 4.5 lists primary options that display parametric density estimates on the histogram. You can specify each primary option once in a given HISTOGRAM statement, and each primary option can display multiple curves from its family on the histogram.

#### Table 4.5 Primary Options for Parametric Fitted Distribution

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA(beta-options)</td>
<td>fits beta distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameters $\alpha$ and $\beta$</td>
</tr>
<tr>
<td>EXPONENTIAL(exponential-options)</td>
<td>fits exponential distribution with threshold parameter $\theta$ and scale parameter $\sigma$</td>
</tr>
<tr>
<td>GAMMA(gamma-options)</td>
<td>fits gamma distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameter $\alpha$</td>
</tr>
<tr>
<td>GUMBEL(Gumbel-options)</td>
<td>fits gumbel distribution with location parameter $\mu$, and scale parameter $\sigma$</td>
</tr>
<tr>
<td>IGAUSS(iGauss-options)</td>
<td>fits inverse Gaussian distribution with location parameter $\mu$, and shape parameter $\lambda$</td>
</tr>
<tr>
<td>LOGNORMAL(lognormal-options)</td>
<td>fits lognormal distribution with threshold parameter $\theta$, scale parameter $\xi$, and shape parameter $\sigma$</td>
</tr>
<tr>
<td>NORMAL(normal-options)</td>
<td>fits normal distribution with mean $\mu$ and standard deviation $\sigma$</td>
</tr>
<tr>
<td>PARETO(Pareto-options)</td>
<td>fits generalized Pareto distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameter $\alpha$</td>
</tr>
<tr>
<td>POWER(power-options)</td>
<td>fits power function distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameter $\alpha$</td>
</tr>
<tr>
<td>RAYLEIGH(Rayleigh-options)</td>
<td>fits Rayleigh distribution with threshold parameter $\theta$, and scale parameter $\sigma$</td>
</tr>
</tbody>
</table>
Table 4.5 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB(S_B-options)</td>
<td>fits Johnson $S_B$ distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameters $\delta$ and $\gamma$</td>
</tr>
<tr>
<td>SU(S_U-options)</td>
<td>fits Johnson $S_U$ distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameters $\delta$ and $\gamma$</td>
</tr>
<tr>
<td>WEIBULL(Weibull-options)</td>
<td>fits Weibull distribution with threshold parameter $\theta$, scale parameter $\sigma$, and shape parameter $c$</td>
</tr>
</tbody>
</table>

Table 4.6 lists secondary options that specify parameters for fitted parametric distributions and that control the display of fitted curves. Specify these secondary options in parentheses after the primary distribution option. For example, you can fit a normal curve by specifying the NORMAL option as follows:

```plaintext
proc univariate;
    histogram / normal(color=red mu=10 sigma=0.5);
run;
```

The COLOR= normal-option draws the curve in red, and the MU= and SIGMA= normal-options specify the parameters $\mu = 10$ and $\sigma = 0.5$ for the curve. Note that the sample mean and sample standard deviation are used to estimate $\mu$ and $\sigma$, respectively, when the MU= and SIGMA= normal-options are not specified.

You can specify lists of values for secondary options to display more than one fitted curve from the same distribution family on a histogram. Option values are matched by list position. You can specify the value EST in a list of distribution parameter values to use an estimate of the parameter.

For example, the following code displays two normal curves on a histogram:

```plaintext
proc univariate;
    histogram / normal(color=(red blue) mu=10 est sigma=0.5 est);
run;
```

The first curve is red, with $\mu = 10$ and $\sigma = 0.5$. The second curve is blue, with $\mu$ equal to the sample mean and $\sigma$ equal to the sample standard deviation.

See the section “Formulas for Fitted Continuous Distributions” on page 401 for detailed information about the families of parametric distributions that you can fit with the HISTOGRAM statement.

Table 4.6 Secondary Options for Parametric Distributions

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR=</td>
<td>specifies colors of density curves</td>
</tr>
<tr>
<td>CONTENTS=</td>
<td>specifies table of contents entry for density curve grouping</td>
</tr>
<tr>
<td>FILL</td>
<td>fills area under density curve</td>
</tr>
<tr>
<td>L=</td>
<td>specifies line types of density curves</td>
</tr>
</tbody>
</table>
### Table 4.6 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIDPERCENTS</td>
<td>prints table of midpoints of histogram intervals</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>suppresses tables summarizing curves</td>
</tr>
<tr>
<td>PERCENTS=</td>
<td>lists percents for which quantiles calculated from data and quantiles estimated from curves are tabulated</td>
</tr>
<tr>
<td>W=</td>
<td>specifies widths of density curves</td>
</tr>
<tr>
<td><strong>Beta-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies first shape parameter $\alpha$ for beta curve</td>
</tr>
<tr>
<td>BETA=</td>
<td>specifies second shape parameter $\beta$ for beta curve</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for beta curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies lower threshold parameter $\theta$ for beta curve</td>
</tr>
<tr>
<td><strong>Exponential-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for exponential curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for exponential curve</td>
</tr>
<tr>
<td><strong>Gamma-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$ for gamma curve</td>
</tr>
<tr>
<td>ALPHADELTA=</td>
<td>specifies change in successive estimates of $\alpha$ at which the Newton-Raphson approximation of $\hat{\alpha}$ terminates</td>
</tr>
<tr>
<td>ALPHAINITIAL=</td>
<td>specifies initial value for $\alpha$ in the Newton-Raphson approximation of $\hat{\alpha}$</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{\alpha}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for gamma curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for gamma curve</td>
</tr>
<tr>
<td><strong>Gumbel-Options</strong></td>
<td></td>
</tr>
<tr>
<td>EDFNSAMPLES=</td>
<td>specifies number of samples for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>EDFSEED=</td>
<td>specifies seed value for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>MU=</td>
<td>specifies location parameter $\mu$ for gumbel curve</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for gumbel curve</td>
</tr>
<tr>
<td><strong>IGauss-Options</strong></td>
<td></td>
</tr>
<tr>
<td>EDFNSAMPLES=</td>
<td>specifies number of samples for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>EDFSEED=</td>
<td>specifies seed value for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>LAMBDA=</td>
<td>specifies shape parameter $\lambda$ for inverse Gaussian curve</td>
</tr>
<tr>
<td>MU=</td>
<td>specifies location parameter $\mu$ for inverse Gaussian curve</td>
</tr>
<tr>
<td><strong>Lognormal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for lognormal curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for lognormal curve</td>
</tr>
<tr>
<td>ZETA=</td>
<td>specifies scale parameter $\zeta$ for lognormal curve</td>
</tr>
<tr>
<td><strong>Normal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies mean $\mu$ for normal curve</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies standard deviation $\sigma$ for normal curve</td>
</tr>
<tr>
<td><strong>Pareto-Options</strong></td>
<td></td>
</tr>
<tr>
<td>EDFNSAMPLES=</td>
<td>specifies number of samples for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>EDFSEED=</td>
<td>specifies seed value for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$ for generalized Pareto curve</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for generalized Pareto curve</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for generalized Pareto curve</td>
</tr>
<tr>
<td><strong>Power-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$ for power function curve</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for power function curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for power function curve</td>
</tr>
<tr>
<td><strong>Rayleigh-Options</strong></td>
<td></td>
</tr>
<tr>
<td>EDFNSAMPLES=</td>
<td>specifies number of samples for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>EDFSEED=</td>
<td>specifies seed value for EDF goodness-of-fit simulation</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for Rayleigh curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for Rayleigh curve</td>
</tr>
<tr>
<td><strong>Johnson S_B-Options</strong></td>
<td></td>
</tr>
<tr>
<td>DELTA=</td>
<td>specifies first shape parameter $\delta$ for Johnson $S_B$ curve</td>
</tr>
<tr>
<td>FITINTERVAL=</td>
<td>specifies z-value for method of percentiles</td>
</tr>
<tr>
<td>FITMETHOD=</td>
<td>specifies method of parameter estimation</td>
</tr>
<tr>
<td>FITTOLERANCE=</td>
<td>specifies tolerance for method of percentiles</td>
</tr>
<tr>
<td>GAMMA=</td>
<td>specifies second shape parameter $\gamma$ for Johnson $S_B$ curve</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for Johnson $S_B$ curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies lower threshold parameter $\theta$ for Johnson $S_B$ curve</td>
</tr>
<tr>
<td><strong>Johnson S_U-Options</strong></td>
<td></td>
</tr>
<tr>
<td>DELTA=</td>
<td>specifies first shape parameter $\delta$ for Johnson $S_U$ curve</td>
</tr>
<tr>
<td>FITINTERVAL=</td>
<td>specifies z-value for method of percentiles</td>
</tr>
<tr>
<td>FITMETHOD=</td>
<td>specifies method of parameter estimation</td>
</tr>
<tr>
<td>FITTOLERANCE=</td>
<td>specifies tolerance for method of percentiles</td>
</tr>
<tr>
<td>GAMMA=</td>
<td>specifies second shape parameter $\gamma$ for Johnson $S_U$ curve</td>
</tr>
<tr>
<td>OPTBOUNDRANGE=</td>
<td>specifies the sampling range for parameter starting values in MLE optimization</td>
</tr>
<tr>
<td>OPTMAXITER=</td>
<td>specifies an iteration limit for MLE optimization</td>
</tr>
<tr>
<td>OPTMAXSTARTS=</td>
<td>specifies the maximum number of starting points to be used for MLE optimization</td>
</tr>
<tr>
<td>OPTPRINT</td>
<td>prints an iteration history for MLE optimization</td>
</tr>
<tr>
<td>OPTSEED=</td>
<td>specifies a seed value for MLE optimization</td>
</tr>
<tr>
<td>OPTTOLERANCE=</td>
<td>specifies the optimality tolerance for MLE optimization</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for Johnson $S_U$ curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies lower threshold parameter $\theta$ for Johnson $S_U$ curve</td>
</tr>
<tr>
<td><strong>Weibull-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies shape parameter $c$ for Weibull curve</td>
</tr>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{c}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$ for Weibull curve</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$ for Weibull curve</td>
</tr>
</tbody>
</table>
Nonparametric Density Estimation Options

Use the option `KERNEL(kernel-options)` to compute kernel density estimates. Specify the following secondary options in parentheses after the KERNEL option to control features of density estimates requested with the KERNEL option.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=</td>
<td>specifies standardized bandwidth parameter $c$</td>
</tr>
<tr>
<td>COLOR=</td>
<td>specifies color of the kernel density curve</td>
</tr>
<tr>
<td>FILL</td>
<td>fills area under kernel density curve</td>
</tr>
<tr>
<td>K=</td>
<td>specifies type of kernel function</td>
</tr>
<tr>
<td>L=</td>
<td>specifies line type used for kernel density curve</td>
</tr>
<tr>
<td>LOWER=</td>
<td>specifies lower bound for kernel density curve</td>
</tr>
<tr>
<td>UPPER=</td>
<td>specifies upper bound for kernel density curve</td>
</tr>
<tr>
<td>W=</td>
<td>specifies line width for kernel density curve</td>
</tr>
</tbody>
</table>

General Options

Table 4.8 summarizes options for enhancing histograms.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BARLABEL=</td>
<td>produces labels above histogram bars</td>
</tr>
<tr>
<td>CLIPCURVES</td>
<td>scales vertical axis without considering fitted curves</td>
</tr>
<tr>
<td>ENDPOINTS=</td>
<td>lists endpoints for histogram intervals</td>
</tr>
<tr>
<td>GRID</td>
<td>creates a grid</td>
</tr>
<tr>
<td>HANGING</td>
<td>constructs hanging histogram</td>
</tr>
<tr>
<td>HREF=</td>
<td>specifies reference lines perpendicular to the horizontal axis</td>
</tr>
<tr>
<td>HREFLABELS=</td>
<td>specifies labels for HREF= lines</td>
</tr>
<tr>
<td>HREFLABPOS=</td>
<td>specifies vertical position of labels for HREF= lines</td>
</tr>
<tr>
<td>MIDPOINTS=</td>
<td>specifies midpoints for histogram intervals</td>
</tr>
<tr>
<td>NENDPOINTS=</td>
<td>specifies number of histogram interval endpoints</td>
</tr>
<tr>
<td>NMIDPOINTS=</td>
<td>specifies number of histogram interval midpoints</td>
</tr>
<tr>
<td>NOBARS</td>
<td>suppresses histogram bars</td>
</tr>
<tr>
<td>NOH LABEL</td>
<td>suppresses label for horizontal axis</td>
</tr>
<tr>
<td>NO PLOT</td>
<td>suppresses plot</td>
</tr>
<tr>
<td>NOV LABEL</td>
<td>suppresses label for vertical axis</td>
</tr>
<tr>
<td>NOV TICK</td>
<td>suppresses tick marks and tick mark labels for vertical axis</td>
</tr>
<tr>
<td>RTINCLUDE</td>
<td>includes right endpoint in interval</td>
</tr>
<tr>
<td>VAXISLABEL=</td>
<td>specifies label for vertical axis</td>
</tr>
<tr>
<td>VREF=</td>
<td>specifies reference lines perpendicular to the vertical axis</td>
</tr>
<tr>
<td>VREFLABELS=</td>
<td>specifies labels for VREF= lines</td>
</tr>
<tr>
<td>VREFLABPOS=</td>
<td>specifies horizontal position of labels for VREF= lines</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>VSCALE=</td>
<td>specifies scale for vertical axis</td>
</tr>
<tr>
<td><strong>Options for Traditional Graphics Output</strong></td>
<td></td>
</tr>
<tr>
<td>ANNOTATE=</td>
<td>specifies annotate data set</td>
</tr>
<tr>
<td>BARWIDTH=</td>
<td>specifies width for the bars</td>
</tr>
<tr>
<td>CAXIS=</td>
<td>specifies color for axis</td>
</tr>
<tr>
<td>CBARLINE=</td>
<td>specifies color for outlines of histogram bars</td>
</tr>
<tr>
<td>CFILL=</td>
<td>specifies color for filling under curve</td>
</tr>
<tr>
<td>CFRAME=</td>
<td>specifies color for frame</td>
</tr>
<tr>
<td>CGRID=</td>
<td>specifies color for grid lines</td>
</tr>
<tr>
<td>CHREF=</td>
<td>specifies colors for HREF= lines</td>
</tr>
<tr>
<td>CLIPREF</td>
<td>draws reference lines behind histogram bars</td>
</tr>
<tr>
<td>CSTATREF=</td>
<td>specifies colors for STATREF= lines</td>
</tr>
<tr>
<td>CTEXT=</td>
<td>specifies color for text</td>
</tr>
<tr>
<td>CVREF=</td>
<td>specifies colors for VREF= lines</td>
</tr>
<tr>
<td>DESCRIPTION=</td>
<td>specifies description for plot in graphics catalog</td>
</tr>
<tr>
<td>FONT=</td>
<td>specifies software font for text</td>
</tr>
<tr>
<td>FRONTREF</td>
<td>draws reference lines in front of histogram bars</td>
</tr>
<tr>
<td>HAXIS=</td>
<td>specifies AXIS statement for horizontal axis</td>
</tr>
<tr>
<td>HEIGHT=</td>
<td>specifies height of text used outside framed areas</td>
</tr>
<tr>
<td>HMINOR=</td>
<td>specifies number of horizontal minor tick mark</td>
</tr>
<tr>
<td>HOFFSET=</td>
<td>specifies offset for horizontal axis</td>
</tr>
<tr>
<td>INFONT=</td>
<td>specifies software font for text inside framed areas</td>
</tr>
<tr>
<td>INHEIGHT=</td>
<td>specifies height of text inside framed areas</td>
</tr>
<tr>
<td>INTERBAR=</td>
<td>specifies space between histogram bars</td>
</tr>
<tr>
<td>LGRID=</td>
<td>specifies a line type for grid lines</td>
</tr>
<tr>
<td>LHREF=</td>
<td>specifies line types for HREF= lines</td>
</tr>
<tr>
<td>LSTATREF=</td>
<td>specifies line types for STATREF= lines</td>
</tr>
<tr>
<td>LVREF=</td>
<td>specifies line types for VREF= lines</td>
</tr>
<tr>
<td>NAME=</td>
<td>specifies name for plot in graphics catalog</td>
</tr>
<tr>
<td>NOFRAME</td>
<td>suppresses frame around plotting area</td>
</tr>
<tr>
<td>PFILL=</td>
<td>specifies pattern for filling under curve</td>
</tr>
<tr>
<td>TURNVLABELS</td>
<td>turns and vertically strings out characters in labels for vertical axis</td>
</tr>
<tr>
<td>VAXIS=</td>
<td>specifies AXIS statement or values for vertical axis</td>
</tr>
<tr>
<td>VMINOR=</td>
<td>specifies number of vertical minor tick mark</td>
</tr>
<tr>
<td>VOFFSET=</td>
<td>specifies length of offset at upper end of vertical axis</td>
</tr>
<tr>
<td>WAXIS=</td>
<td>specifies line thickness for axes and frame</td>
</tr>
<tr>
<td>WBARLINE=</td>
<td>specifies line thickness for bar outlines</td>
</tr>
<tr>
<td>WGRID=</td>
<td>specifies line thickness for grid</td>
</tr>
<tr>
<td><strong>Options for ODS Graphics Output</strong></td>
<td></td>
</tr>
<tr>
<td>ODSFOOTNOTE=</td>
<td>specifies footnote displayed on histogram</td>
</tr>
<tr>
<td>ODSFOOTNOTE2=</td>
<td>specifies secondary footnote displayed on histogram</td>
</tr>
<tr>
<td>ODSTITLE=</td>
<td>specifies title displayed on histogram</td>
</tr>
<tr>
<td>ODSTITLE2=</td>
<td>specifies secondary title displayed on histogram</td>
</tr>
<tr>
<td>OVERLAY</td>
<td>overlays histograms for different class levels</td>
</tr>
</tbody>
</table>
Table 4.8  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Options for Comparative Plots</strong></td>
<td></td>
</tr>
<tr>
<td>ANNOKEY</td>
<td>applies annotation requested in ANNOTATE= data set to key cell only</td>
</tr>
<tr>
<td>CFRAMESIDE=</td>
<td>specifies color for filling frame for row labels</td>
</tr>
<tr>
<td>CFRAMETOP=</td>
<td>specifies color for filling frame for column labels</td>
</tr>
<tr>
<td>CPROP=</td>
<td>specifies color for proportion of frequency bar</td>
</tr>
<tr>
<td>CTEXTSIDE=</td>
<td>specifies color for row labels of comparative histograms</td>
</tr>
<tr>
<td>CTEXTTOP=</td>
<td>specifies color for column labels of comparative histograms</td>
</tr>
<tr>
<td>INTERTILE=</td>
<td>specifies distance between tiles</td>
</tr>
<tr>
<td>MAXNBIN=</td>
<td>specifies maximum number of bins to display</td>
</tr>
<tr>
<td>MAXSIGMAs=</td>
<td>limits the number of bins that display to within a specified number of</td>
</tr>
<tr>
<td></td>
<td>standard deviations above and below mean of data in key cell</td>
</tr>
<tr>
<td>NCOLS=</td>
<td>specifies number of columns in comparative histogram</td>
</tr>
<tr>
<td>NROWS=</td>
<td>specifies number of rows in comparative histogram</td>
</tr>
<tr>
<td><strong>Miscellaneous Options</strong></td>
<td></td>
</tr>
<tr>
<td>CONTENTS=</td>
<td>specifies table of contents entry for histogram grouping</td>
</tr>
<tr>
<td>MIDPERCENTS</td>
<td>creates table of histogram intervals</td>
</tr>
<tr>
<td>NOTABCONTENTS</td>
<td>suppresses table of contents entries for tables produced by HISTOGRAM</td>
</tr>
<tr>
<td></td>
<td>statement</td>
</tr>
<tr>
<td>OUTHISTOGRAM=</td>
<td>creates a data set containing information about histogram intervals</td>
</tr>
<tr>
<td>OUTKERNEL=</td>
<td>creates a data set containing kernel density estimates</td>
</tr>
</tbody>
</table>

**Dictionary of Options**

The following entries provide detailed descriptions of options in the HISTOGRAM statement. Options marked with † are applicable only when traditional graphics are produced. See the section “Dictionary of Common Options” on page 369 for detailed descriptions of options common to all plot statements.

**ALPHA=value-list**

specifies the shape parameter $\alpha$ for fitted curves requested with the BETA, GAMMA, PARETO and POWER options. Enclose the ALPHA= option in parentheses after the distribution keyword. By default, or if you specify the value EST, the procedure calculates a maximum likelihood estimate for $\alpha$. You can specify A= as an alias for ALPHA= if you use it as a **beta-option**. You can specify SHAPE= as an alias for ALPHA= if you use it as a **gamma-option**.

**BARLABEL=COUNT | PERCENT | PROPORTION**

displays labels above the histogram bars. If you specify BARLABEL=COUNT, the label shows the number of observations associated with a given bar. If you specify BARLABEL=PERCENT, the label shows the percentage of observations represented by that bar. If you specify BARLABEL=PROPORTION, the label displays the proportion of observations associated with the bar.

**BARWIDTH=value**

specifies the width of the histogram bars in percentage screen units. If both the BARWIDTH= and INTERBAR= options are specified, the INTERBAR= option takes precedence.
HISTOGRAM Statement

BETA < (beta-options)>
displays fitted beta density curves on the histogram. The BETA option can occur only once in a
HISTOGRAM statement, but it can request any number of beta curves. The beta distribution is
bounded below by the parameter $\theta$ and above by the value $\theta + \sigma$. Use the THETA= and SIGMA=
beta-options to specify these parameters. By default, THETA=0 and SIGMA=1. You can specify
THETA=EST and SIGMA=EST to request maximum likelihood estimates for $\theta$ and $\sigma$.

The beta distribution has two shape parameters: $\alpha$ and $\beta$. If these parameters are known, you can
specify their values with the ALPHA= and BETA= beta-options. By default, the procedure computes
maximum likelihood estimates for $\alpha$ and $\beta$. **NOTE:** Three- and four-parameter maximum likelihood
estimation may not always converge.

Table 4.6 lists secondary options you can specify with the BETA option. See the section “Beta
Distribution” on page 401 for details and Example 4.21 for an example that uses the BETA option.

BETA=value-list
B=value-list
specifies the second shape parameter $\beta$ for beta density curves requested with the BETA option.
Enclose the BETA= option in parentheses after the BETA option. By default, or if you specify the
value EST, the procedure calculates a maximum likelihood estimate for $\beta$.

C=value-list
specifies the shape parameter $c$ for Weibull density curves requested with the WEIBULL option.
Enclose the C= Weibull-option in parentheses after the WEIBULL option. By default, or if you specify the
value EST, the procedure calculates a maximum likelihood estimate for $c$. You can specify the
SHAPE= Weibull-option as an alias for the C= Weibull-option.

C=value-list
specifies the standardized bandwidth parameter $c$ for kernel density estimates requested with the
KERNEL option. Enclose the C= kernel-option in parentheses after the KERNEL option. You can
specify a list of values to request multiple estimates. You can specify the value MISE to produce the
estimate with a bandwidth that minimizes the approximate mean integrated square error (MISE), or
SJPI to select the bandwidth by using the Sheather-Jones plug-in method.

You can also use the C= kernel-option with the K= kernel-option (which specifies the kernel function) to
compute multiple estimates. If you specify more kernel functions than bandwidths, the last bandwidth
in the list is repeated for the remaining estimates. Similarly, if you specify more bandwidths than
kernel functions, the last kernel function is repeated for the remaining estimates. If you do not specify
the C= kernel-option, the bandwidth that minimizes the approximate MISE is used for all the estimates.

See the section “Kernel Density Estimates” on page 418 for more information about kernel density
estimates.

† CBARLINE=color
specifies the color for the outline of the histogram bars when producing traditional graphics. The
option does not apply to ODS Graphics output.

† 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) when producing traditional graphics. See the entries for the FILL and PFILL=
options for additional details. Refer to **SAS/GRAPH: Reference** for a list of colors. The option does not
apply to ODS Graphics output.
† CGRID=color
specifies the color for grid lines when a grid displays on the histogram in traditional graphics. This option also produces a grid if the GRID= option is not specified.

CLIPCURVES
scales the vertical axis without taking fitted curves into consideration. Curves that extend above the tallest histogram bar can be clipped. You can use this option to avoid compression of the histogram bars due to extremely high fitted curve peaks.

† CLIPREF
draws reference lines requested with the HREF= and VREF= options behind the histogram bars. When the GSTYLE system option is in effect for traditional graphics, reference lines are drawn in front of the bars by default.

CONTENTS=
specifies the table of contents grouping entry for tables associated with a density curve. Enclose the CONTENTS= option in parentheses after the distribution option. You can specify CONTENTS="" to suppress the grouping entry.

DELTA=value-list
specifies the first shape parameter \( \delta \) for Johnson SB and Johnson SU distribution functions requested with the SB and SU options. Enclose the DELTA= option in parentheses after the SB or SU option. If you do not specify a value for \( \delta \), or if you specify the value EST, the procedure calculates an estimate.

EDFNAMES=value
specifies the number of simulation samples used to compute \( p \)-values for EDF goodness-of-fit statistics for density curves requested with the GUMBEL, IGAUSS, PARETO, and RAYLEIGH options. Enclose the EDFNAMES= option in parentheses after the distribution option. The default value is 500.

EDFSEED=value
specifies an integer value used to start the pseudo-random number generator when creating simulation samples for computing EDF goodness-of-fit statistic \( p \)-values for density curves requested with the GUMBEL, IGAUSS, PARETO, and RAYLEIGH options. Enclose the EDFSEED= option in parentheses after the distribution option. By default, the procedure uses a random number seed generated from reading the time of day from the computer's clock.

ENDPOINTS < =values | KEY | UNIFORM >
uses histogram bin endpoints as the tick mark values for the horizontal axis and determines how to compute the bin width of the histogram bars. The values specify both the left and right endpoint of each histogram interval. The width of the histogram bars is the difference between consecutive endpoints. The procedure uses the same values for all variables.
The range of endpoints must cover the range of the data. For example, if you specify

\[ \text{endpoints}=2 \text{ to } 10 \text{ by 2} \]

then all of the observations must fall in the intervals \([2,4) \ [4,6) \ [6,8) \ [8,10]\). You also must use evenly spaced endpoints which you list in increasing order.

**KEY** determines the endpoints for the data in the key cell. The initial number of endpoints is based on the number of observations in the key cell by using the method of Terrell and Scott (1985). The procedure extends the endpoint list for the key cell in either direction as necessary until it spans the data in the remaining cells.

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

Neither **KEY** nor **UNIFORM** apply unless you use the CLASS statement.

If you omit **ENDPOINTS**, the procedure uses the histogram midpoints as horizontal axis tick values. If you specify **ENDPOINTS**, the procedure computes the endpoints by using an algorithm (Terrell and Scott 1985) that is primarily applicable to continuous data that are approximately normally distributed.

If you specify both **MIDPOINTS=** and **ENDPOINTS**, the procedure issues a warning message and uses the endpoints.

If you specify **RTINCLUDE**, the procedure includes the right endpoint of each histogram interval in that interval instead of including the left endpoint.

If you use a CLASS statement and specify **ENDPOINTS**, the procedure uses **ENDPOINTS=KEY** as the default. However if the key cell is empty, then the procedure uses **ENDPOINTS=UNIFORM**.

**EXPONENTIAL** <(exponential-options)> 
**EXP** <(exponential-options)>
displays fitted exponential density curves on the histogram. The **EXPONENTIAL** option can occur only once in a HISTOGRAM statement, but it can request any number of exponential curves. The parameter \( \theta \) must be less than or equal to the minimum data value. Use the THETA= exponential-option to specify \( \theta \). By default, THETA=0. You can specify THETA=EST to request the maximum likelihood estimate for \( \theta \). Use the SIGMA= exponential-option to specify \( \sigma \). By default, the procedure computes a maximum likelihood estimate for \( \sigma \). Table 4.6 lists options you can specify with the **EXPONENTIAL** option. See the section “Exponential Distribution” on page 403 for details.
FILL
fills areas under the fitted density curve or the kernel density estimate with colors and patterns. The FILL option can occur with only one fitted curve. Enclose the FILL option in parentheses after a density curve option or the KERNEL option. The CFILL= and PFILL= options specify the color and pattern for the area under the curve when producing traditional graphics. For a list of available colors and patterns, see SAS/GRAPH: Reference.

† FRONTREF
draws reference lines requested with the HREF= and VREF= options in front of the histogram bars. When the NOGSTYLE system option is in effect for traditional graphics, reference lines are drawn behind the histogram bars by default, and they can be obscured by filled bars.

GAMMA < (gamma-options) >
displays fitted gamma density curves on the histogram. The GAMMA option can occur only once in a HISTOGRAM statement, but it can request any number of gamma curves. The parameter \( \theta \) must be less than the minimum data value. Use the THETA= gamma-option to specify \( \theta \). By default, THETA=0. You can specify THETA=EST to request the maximum likelihood estimate for \( \theta \). Use the ALPHA= and the SIGMA= gamma-options to specify the shape parameter \( \alpha \) and the scale parameter \( \sigma \). By default, PROC UNIVARIATE computes maximum likelihood estimates for \( \alpha \) and \( \sigma \). The procedure calculates the maximum likelihood estimate of \( \alpha \) iteratively by using the Newton-Raphson approximation. Table 4.6 lists options you can specify with the GAMMA option. See the section “Gamma Distribution” on page 403 for details, and see Example 4.22 for an example that uses the GAMMA option.

GAMMA=value-list
specifies the second shape parameter \( \gamma \) for Johnson SB and Johnson SU distribution functions requested with the SB and SU options. Enclose the GAMMA= option in parentheses after the SB or SU option. If you do not specify a value for \( \gamma \), or if you specify the value EST, the procedure calculates an estimate.

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

GUMBEL < (Gumbel-options) >
displays fitted Gumbel density curves on the histogram. The GUMBEL option can occur only once in a HISTOGRAM statement, but it can request any number of Gumbel curves. Use the MU= and the SIGMA= Gumbel-options to specify the location parameter \( \mu \) and the scale parameter \( \sigma \). By default, PROC UNIVARIATE computes maximum likelihood estimates for \( \mu \) and \( \sigma \). Table 4.6 lists options you can specify with the GUMBEL option. See the section “Gumbel Distribution” on page 404 for details about the Gumbel distribution.

HANGING

HANG
requests a hanging histogram, as illustrated in Figure 4.7.
You can use the HANGING option only when exactly one fitted density curve is requested. A hanging histogram aligns the tops of the histogram bars (displayed as lines) with the fitted curve. The lines are positioned at the midpoints of the histogram bins. A hanging histogram is a goodness-of-fit diagnostic in the sense that the closer the lines are to the horizontal axis, the better the fit. Hanging histograms are discussed by Tukey (1977), Wainer (1974), and Velleman and Hoaglin (1981).

† HOFFSET=value

specifies the offset, in percentage screen units, at both ends of the horizontal axis. You can use HOFFSET=0 to eliminate the default offset.

IGAUSS < (iGauss-options) >

displays fitted inverse Gaussian density curves on the histogram. The IGAUSS option can occur only once in a HISTOGRAM statement, but it can request any number of inverse Gaussian curves. Use the MU= and the LAMBDA= iGauss-options to specify the location parameter $\mu$ and the shape parameter $\lambda$. By default, PROC UNIVARIATE uses the sample mean for $\mu$ and computes a maximum likelihood estimate for $\lambda$. Table 4.6 lists options you can specify with the IGAUSS option. See the section “Inverse Gaussian Distribution” on page 405 for details.
† INTERBAR=\textit{value}

specifies the space between histogram bars in percentage screen units. If both the INTERBAR= and BARWIDTH= options are specified, the INTERBAR= option takes precedence.

**K=NORMAL | QUADRATIC | TRIANGULAR**

specifies the kernel function (normal, quadratic, or triangular) used to compute a kernel density estimate. You can specify a list of values to request multiple estimates. You must enclose this option in parentheses after the KERNEL option. You can also use the \textit{K= kernel-option} with the \textit{C= kernel-option}, which specifies standardized bandwidths. If you specify more kernel functions than bandwidths, the procedure repeats the last bandwidth in the list for the remaining estimates. Similarly, if you specify more bandwidths than kernel functions, the procedure repeats the last kernel function for the remaining estimates. By default, K=NORMAL.

KERNEL\textless (kernel-options)\textgreater

superimposes kernel density estimates on the histogram. By default, the procedure uses the AMISE method to compute kernel density estimates. To request multiple kernel density estimates on the same histogram, specify a list of values for the \textit{C= kernel-option} or \textit{K= kernel-option}. Table 4.7 lists options you can specify with the KERNEL option. See the section “Kernel Density Estimates” on page 418 for more information about kernel density estimates, and see Example 4.23.

LAMBDA=\textit{value}

specifies the shape parameter \( \lambda \) for fitted curves requested with the IGAUSS option. Enclose the LAMBDA= option in parentheses after the IGAUSS distribution keyword. If you do not specify a value for \( \lambda \), the procedure calculates a maximum likelihood estimate.

† LGRID=linetype

specifies the line type for the grid when a grid displays on the histogram. This option also creates a grid if the GRID option is not specified.

LOGNORMAL\textless (lognormal-options)\textgreater

displays fitted lognormal density curves on the histogram. The LOGNORMAL option can occur only once in a HISTOGRAM statement, but it can request any number of lognormal curves. The parameter \( \theta \) must be less than the minimum data value. Use the THETA= \textit{lognormal-option} to specify \( \theta \). By default, THETA=0. You can specify THETA=EST to request the maximum likelihood estimate for \( \theta \). Use the SIGMA= and ZETA= \textit{lognormal-options} to specify \( \sigma \) and \( \zeta \). By default, the procedure computes maximum likelihood estimates for \( \sigma \) and \( \zeta \). Table 4.6 lists options you can specify with the LOGNORMAL option. See the section “Lognormal Distribution” on page 406 for details, and see Example 4.22 and Example 4.24 for examples using the LOGNORMAL option.

LOWER=\textit{value-list}

specifies lower bounds for kernel density estimates requested with the KERNEL option. Enclose the LOWER= option in parentheses after the KERNEL option. If you specify more kernel estimates than lower bounds, the last lower bound is repeated for the remaining estimates. The default is a missing value, indicating no lower bounds for fitted kernel density curves.

MAXNBIN=\textit{n}

limits the number of bins displayed in the comparative histogram. This option is useful when the scales or ranges of the data distributions differ greatly from cell to cell. By default, the bin size and midpoints are determined for the key cell, and then the midpoint list is extended to accommodate the data ranges for the remaining cells. However, if the cell scales differ considerably, the resulting number of bins can
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. This option is not available unless you specify the CLASS statement. The MAXNBIN= option is an alternative to the MAXSIGMAS= option.

**MAXSIGMAS=** *value*

limits the number of bins displayed in the comparative histogram 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, the bin size and midpoints are determined for the key cell, and then the midpoint list is extended to accommodate the data ranges for the remaining cells. However, if the cell scales differ considerably, the resulting number of bins can 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. This option is not available unless you specify the CLASS statement.

**MIDPERCENTS**

requests a table listing the midpoints and percentage of observations in each histogram interval. If you specify MIDPERCENTS in parentheses after a density estimate option, the procedure 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). See Example 4.18.

**MIDPOINTS=** *values | KEY | UNIFORM*

specifies how to determine the midpoints for the histogram intervals, where *values* determines the width of the histogram bars as the difference between consecutive midpoints. The procedure uses the same values for all variables.

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. You must use evenly spaced midpoints listed 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). The procedure extends the midpoint list for the key cell in either direction as necessary until it spans the data in the remaining cells.

**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).

Neither KEY nor UNIFORM apply unless you use the CLASS statement. By default, if you use a CLASS statement, MIDPOINTS=KEY; however, if the key cell is empty then MIDPOINTS=UNIFORM. Otherwise, the procedure computes the midpoints by using an algorithm (Terrell and Scott 1985) that is primarily applicable to continuous data that are approximately normally distributed.
MU=value-list
specifies the parameter $\mu$ for Gumbel, inverse Gaussian, and normal density curves requested with the GUMBEL, IGAUSS, and NORMAL options, respectively. Enclose the MU= option in parentheses after the distribution keyword. By default, or if you specify the value EST, the procedure uses the sample mean for $\mu$ for normal and inverse Gaussian distributions and computes a maximum likelihood estimate of $\mu$ for the Gumbel distribution. For more detail please see “Inverse Gaussian Distribution” on page 405 and “Gumbel Distribution” on page 404.

NENDPOINTS=n
uses histogram interval endpoints as the tick mark values for the horizontal axis and determines the number of bins.

NMIDPOINTS=n
specifies the number of histogram intervals.

NOBARS
suppresses drawing of histogram bars, which is useful for viewing fitted curves only.

NOPLOT
NOCHART
suppresses the creation of a plot. Use this option when you only want to tabulate summary statistics for a fitted density or create an OUTHISTOGRAM= data set.

NOPRINT
suppresses tables summarizing the fitted curve. Enclose the NOPRINT option in parentheses following the distribution option.

NORMAL< (normal-options) >
displays fitted normal density curves on the histogram. The NORMAL option can occur only once in a HISTOGRAM statement, but it can request any number of normal curves. Use the MU= and SIGMA= normal-options to specify $\mu$ and $\sigma$. By default, the procedure uses the sample mean and sample standard deviation for $\mu$ and $\sigma$. Table 4.6 lists options you can specify with the NORMAL option. See the section “Normal Distribution” on page 406 for details, and see Example 4.19 for an example that uses the NORMAL option.

NOTABCONTENTS
suppresses the table of contents entries for tables produced by the HISTOGRAM statement.

OPTBOUND_RANGE=value
defines the sampling range for each parameter during maximum likelihood estimation for the Johnson $S_U$ distribution. PROC UNIVARIATE computes initial estimates for each parameter by using the method of percentiles. The value determines the range of parameter values around the initial estimate that can be sampled for local optimization starting values. The default is 100.

OPTMAXITER=value
limits the number of iterations that are used by the optimizer in maximum likelihood estimation for the Johnson $S_U$ distribution. The default is 500.
OPTMAXSTARTS=N
defines the maximum number of starting points to be used for local optimization in maximum likelihood estimation for the Johnson \( S_U \) distribution. That is, no more than \( N \) local optimizations are used in the multistart algorithm. The default value is 100.

OPTPRINT
prints the iteration history for the Johnson \( S_U \) distribution maximum likelihood estimation.

OPTSEED=value
specifies a positive integer seed for generating random number sequences in Johnson \( S_U \) distribution maximum likelihood estimation. You can use this option to replicate results from different runs.

OPTTOLERANCE=value
specifies the tolerance for declaring optimality in maximum likelihood estimation for the Johnson \( S_U \) distribution. The default value is 1E–8.

OUTHISTOGRAM=SAS-data-set
OUTHIST=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 (or the lower endpoints of the intervals if you specify the ENDPOINTS option), 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).

PARETO < (Pareto-options) >
displays fitted generalized Pareto density curves on the histogram. The PARETO option can occur only once in a HISTOGRAM statement, but it can request any number of generalized Pareto curves. The parameter \( \theta \) must be less than the minimum data value. Use the THETA= Pareto-option to specify \( \theta \). By default, THETA=0. Use the SIGMA= and the ALPHA= Pareto-options to specify the scale parameter \( \sigma \) and the shape parameter \( \alpha \). By default, PROC UNIVARIATE computes maximum likelihood estimates for \( \sigma \) and \( \alpha \). Table 4.6 lists options you can specify with the PARETO option. See the section “Generalized Pareto Distribution” on page 407 for details.

PERCENTS=values
PERCENT=values
specifies a list of percents for which quantiles calculated from the data and quantiles estimated from the fitted curve are tabulated. The percents must be between 0 and 100. Enclose the PERCENTS= option in parentheses after the curve option. The default percents are 1, 5, 10, 25, 50, 75, 90, 95, and 99.

† PFILL=pattern
specifies a pattern used to fill the bars of the histograms (or the areas under a fitted curve if you also specify the FILL option) when producing traditional graphics. See the entries for the CFILL= and FILL options for additional details. Refer to SAS/GRAPH: Reference for a list of pattern values. The option does not apply to ODS Graphics output.

POWER < (power-options) >
displays fitted power function density curves on the histogram. The POWER option can occur only once in a HISTOGRAM statement, but it can request any number of power function curves. The parameter \( \theta \) must be less than the minimum data value. Use the THETA= and SIGMA= power-options to specify \( \theta \) and \( \sigma \). The default values are 0 and 1, respectively. Use the ALPHA= power-option
to specify the and the shape parameter, $\alpha$. By default, PROC UNIVARIATE computes a maximum likelihood estimate for $\alpha$. Table 4.6 lists options you can specify with the POWER option. See the section “Power Function Distribution” on page 408 for details.

**RAYLEIGH <(Rayleigh-options) >**
displays fitted Rayleigh density curves on the histogram. The RAYLEIGH option can occur only once in a HISTOGRAM statement, but it can request any number of Rayleigh curves. The parameter $\theta$ must be less than the minimum data value. Use the THETA= Rayleigh-option to specify $\theta$. By default, THETA=0. Use the SIGMA= Rayleigh-option to specify the scale parameter $\sigma$. By default, PROC UNIVARIATE computes maximum likelihood estimates for $\sigma$. Table 4.6 lists options you can specify with the RAYLEIGH option. See the section “Rayleigh Distribution” on page 409 for details.

**RTINCLUDE**
includes the right endpoint of each histogram interval in that interval. By default, the left endpoint is included in the histogram interval.

**SB<(SB-options)>**
displays fitted Johnson $S_B$ density curves on the histogram. The SB option can occur only once in a HISTOGRAM statement, but it can request any number of Johnson $S_B$ curves. Use the THETA= and SIGMA= normal-options to specify $\theta$ and $\sigma$. By default, the procedure computes maximum likelihood estimates of $\theta$ and $\sigma$. Table 4.6 lists options you can specify with the SB option. See the section “Johnson $S_B$ Distribution” on page 410 for details.

**SIGMA=value-list**
specifies the parameter $\sigma$ for the fitted density curve when you request the BETA, EXPONENTIAL, GAMMA, GUMBEL, LOGNORMAL, NORMAL, PARETO, POWER, RAYLEIGH, SB, SU, or WEIBULL options.

See Table 4.9 for a summary of how to use the SIGMA= option. You must enclose this option in parentheses after the density curve option. You can specify the value EST to request a maximum likelihood estimate for $\sigma$.

<table>
<thead>
<tr>
<th>Distribution Keyword</th>
<th>SIGMA= Specifies</th>
<th>Default Value</th>
<th>Alias</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA</td>
<td>scale parameter $\sigma$</td>
<td>1</td>
<td>SCALE=</td>
</tr>
<tr>
<td>EXPONENTIAL</td>
<td>scale parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td>SCALE=</td>
</tr>
<tr>
<td>GAMMA</td>
<td>scale parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td>SCALE=</td>
</tr>
<tr>
<td>GUMBEL</td>
<td>scale parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
<tr>
<td>LOGNORMAL</td>
<td>shape parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td>SHAPE=</td>
</tr>
<tr>
<td>NORMAL</td>
<td>scale parameter $\sigma$</td>
<td>standard deviation</td>
<td></td>
</tr>
<tr>
<td>PARETO</td>
<td>scale parameter $\sigma$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>POWER</td>
<td>scale parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
<tr>
<td>RAYLEIGH</td>
<td>scale parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
<tr>
<td>SB</td>
<td>scale parameter $\sigma$</td>
<td>1</td>
<td>SCALE=</td>
</tr>
<tr>
<td>SU</td>
<td>scale parameter $\sigma$</td>
<td>percentile-based estimate</td>
<td></td>
</tr>
<tr>
<td>WEIBULL</td>
<td>scale parameter $\sigma$</td>
<td>maximum likelihood estimate</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.9 Uses of the SIGMA= Option
**HISTOGRAM Statement**

The `SU< (SU-options) >` option displays fitted Johnson S_U density curves on the histogram. The SU option can occur only once in a HISTOGRAM statement, but it can request any number of Johnson S_U curves. Use the THETA= and SIGMA= normal-options to specify \( \theta \) and \( \sigma \). By default, the procedure computes maximum likelihood estimates of \( \theta \) and \( \sigma \). Table 4.6 lists options you can specify with the SU option. See the section “Johnson S_U Distribution” on page 411 for details.

**THETA=value-list**

**THRESHOLD=value-list**

specifies the lower threshold parameter \( \theta \) for curves requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, SB, SU, and WEIBULL options. Enclose the THETA= option in parentheses after the curve option. By default, THETA=0. If you specify the value EST, an estimate is computed for \( \theta \).

**UPPER=value-list**

specifies upper bounds for kernel density estimates requested with the KERNEL option. Enclose the UPPER= option in parentheses after the KERNEL option. If you specify more kernel estimates than upper bounds, the last upper bound is repeated for the remaining estimates. The default is a missing value, indicating no upper bounds for fitted kernel density curves.

**† VOFFSET=value**

specifies the offset, in percentage screen units, at the upper end of the vertical axis.

**VSChALE=COUNT | PERCENT | PROPORTION**

specifies the scale of the vertical axis for a histogram. The value COUNT requests the data be scaled in units of the number of observations per data unit. The value PERCENT requests the data be scaled in units of percent of observations per data unit. The value PROPORTION requests the data be scaled in units of proportion of observations per data unit. The default is PERCENT.

**† WBARLINE=n**

specifies the width of bar outlines when producing traditional graphics. The option does not apply to ODS Graphics output.

**WEIBULL< (Weibull-options) >**

displays fitted Weibull density curves on the histogram. The WEIBULL option can occur only once in a HISTOGRAM statement, but it can request any number of Weibull curves. The parameter \( \theta \) must be less than the minimum data value. Use the THETA= Weibull-option to specify \( \theta \). By default, THETA=0. You can specify THETA=EST to request the maximum likelihood estimate for \( \theta \). Use the C= and SIGMA= Weibull-options to specify the shape parameter \( c \) and the scale parameter \( \sigma \). By default, the procedure computes the maximum likelihood estimates for \( c \) and \( \sigma \). Table 4.6 lists options you can specify with the WEIBULL option. See the section “Weibull Distribution” on page 413 for details, and see Example 4.22 for an example that uses the WEIBULL option.

PROC UNIVARIATE calculates the maximum likelihood estimate of \( a \) iteratively by using the Newton-Raphson approximation. See also the C=, SIGMA=, and THETA= Weibull-options.

**† WGRID=n**

specifies the line thickness for the grid when producing traditional graphics. The option does not apply to ODS Graphics output.
ZETA= value-list

specifies a value for the scale parameter $\zeta$ for lognormal density curves requested with the LOGNORMAL option. Enclose the ZETA= lognormal-option in parentheses after the LOGNORMAL option. By default, or if you specify the value EST, the procedure calculates a maximum likelihood estimate for $\zeta$. You can specify the SCALE= option as an alias for the ZETA= option.

ID Statement

ID variables ;

The ID statement 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 Example 4.3.

You can also include ID variables in the output data set created by an OUTPUT statement by specifying the IDOUT option in the PROC UNIVARIATE statement.

INSET Statement

INSET keywords </options> ;

An INSET statement places a box or table of summary statistics, called an inset, directly in a graph created with a CDFPLOT, HISTOGRAM, PPLOT, PROB PLOT, or QQPLOT statement. The INSET statement must follow the plot statement that creates the plot that you want to augment. The inset appears in all the graphs that the preceding plot statement produces.

You can use multiple INSET statements after a plot statement to add more than one inset to a plot. See Example 4.17.

In an INSET statement, you specify one or more keywords that identify the information to display in the inset. The information is displayed in the order that you request the keywords. Keywords can be any of the following:

- statistical keywords
- primary keywords
- secondary keywords

Statistical Keywords

The available statistical keywords are listed in Table 4.10.
### Table 4.10  Statistical Keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive Statistic Keywords</strong></td>
<td></td>
</tr>
<tr>
<td>CSS</td>
<td>Corrected sum of squares</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of variation</td>
</tr>
<tr>
<td>KURTOSIS</td>
<td>KURT</td>
</tr>
<tr>
<td>MAX</td>
<td>Largest value</td>
</tr>
<tr>
<td>MEAN</td>
<td>Sample mean</td>
</tr>
<tr>
<td>MIN</td>
<td>Smallest value</td>
</tr>
<tr>
<td>MODE</td>
<td>Most frequent value</td>
</tr>
<tr>
<td>N</td>
<td>Sample size</td>
</tr>
<tr>
<td>NEXCL</td>
<td>Number of observations excluded by MAXNBIN= or MAXSIGMAS= option</td>
</tr>
<tr>
<td>NMISS</td>
<td>Number of missing values</td>
</tr>
<tr>
<td>NOBS</td>
<td>Number of observations</td>
</tr>
<tr>
<td>RANGE</td>
<td>Range</td>
</tr>
<tr>
<td>SKEWNESS</td>
<td>SKEW</td>
</tr>
<tr>
<td>STD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>STDMEAN</td>
<td>STDERR</td>
</tr>
<tr>
<td>SUM</td>
<td>Sum of the observations</td>
</tr>
<tr>
<td>SUMWGT</td>
<td>Sum of the weights</td>
</tr>
<tr>
<td>USS</td>
<td>Uncorrected sum of squares</td>
</tr>
<tr>
<td>VAR</td>
<td>Variance</td>
</tr>
<tr>
<td><strong>Percentile Statistic Keywords</strong></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>1st percentile</td>
</tr>
<tr>
<td>P5</td>
<td>5th percentile</td>
</tr>
<tr>
<td>P10</td>
<td>10th percentile</td>
</tr>
<tr>
<td>Q1</td>
<td></td>
</tr>
<tr>
<td>P25</td>
<td>Lower quartile (25th percentile)</td>
</tr>
<tr>
<td>MEDIAN</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td></td>
</tr>
<tr>
<td>P50</td>
<td>Median (50th percentile)</td>
</tr>
<tr>
<td>Q3</td>
<td></td>
</tr>
<tr>
<td>P75</td>
<td>Upper quartile (75th percentile)</td>
</tr>
<tr>
<td>P90</td>
<td>90th percentile</td>
</tr>
<tr>
<td>P95</td>
<td>95th percentile</td>
</tr>
<tr>
<td>P99</td>
<td>99th percentile</td>
</tr>
<tr>
<td>QRANGE</td>
<td>Interquartile range (Q3–Q1)</td>
</tr>
<tr>
<td><strong>Keywords for Distribution-Free Confidence Limits for Percentiles (CIPCTLDF Option)</strong></td>
<td></td>
</tr>
<tr>
<td>P1_LCL_DF</td>
<td>1st percentile lower confidence limit</td>
</tr>
<tr>
<td>P1_UCL_DF</td>
<td>1st percentile upper confidence limit</td>
</tr>
<tr>
<td>P5_LCL_DF</td>
<td>5th percentile lower confidence limit</td>
</tr>
<tr>
<td>P5_UCL_DF</td>
<td>5th percentile upper confidence limit</td>
</tr>
<tr>
<td>P10_LCL_DF</td>
<td>10th percentile lower confidence limit</td>
</tr>
</tbody>
</table>
### Table 4.10 continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10_UCL_DF</td>
<td>10th percentile upper confidence limit</td>
</tr>
<tr>
<td>Q1_LCL_DF</td>
<td></td>
</tr>
<tr>
<td>P25_LCL_DF</td>
<td>Lower quartile (25th percentile) lower confidence limit</td>
</tr>
<tr>
<td>Q1_UCL_DF</td>
<td>Lower quartile (25th percentile) upper confidence limit</td>
</tr>
<tr>
<td>P25_UCL_DF</td>
<td>Lower quartile (25th percentile) upper confidence limit</td>
</tr>
<tr>
<td>MEDIAN_LCL_DF</td>
<td>Median (50th percentile) lower confidence limit</td>
</tr>
<tr>
<td>Q2_LCL_DF</td>
<td>Median (50th percentile) upper confidence limit</td>
</tr>
<tr>
<td>P50_LCL_DF</td>
<td></td>
</tr>
<tr>
<td>MEDIAN_UCL_DF</td>
<td></td>
</tr>
<tr>
<td>Q2_UCL_DF</td>
<td></td>
</tr>
<tr>
<td>P50_UCL_DF</td>
<td>Median (50th percentile) upper confidence limit</td>
</tr>
<tr>
<td>Q3_LCL_DF</td>
<td></td>
</tr>
<tr>
<td>P75_LCL_DF</td>
<td>Upper quartile (75th percentile) lower confidence limit</td>
</tr>
<tr>
<td>Q3_UCL_DF</td>
<td>Upper quartile (75th percentile) upper confidence limit</td>
</tr>
<tr>
<td>P75_UCL_DF</td>
<td>Upper quartile (75th percentile) upper confidence limit</td>
</tr>
<tr>
<td>P90_LCL_DF</td>
<td>90th percentile lower confidence limit</td>
</tr>
<tr>
<td>P90_UCL_DF</td>
<td>90th percentile upper confidence limit</td>
</tr>
<tr>
<td>P95_LCL_DF</td>
<td>95th percentile lower confidence limit</td>
</tr>
<tr>
<td>P95_UCL_DF</td>
<td>95th percentile upper confidence limit</td>
</tr>
<tr>
<td>P99_LCL_DF</td>
<td>99th percentile lower confidence limit</td>
</tr>
<tr>
<td>P99_UCL_DF</td>
<td>99th percentile upper confidence limit</td>
</tr>
</tbody>
</table>

**Keywords Percentile Confidence Limits Assuming Normality (CIPCTLNORMAL Option)**

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1_LCL</td>
<td>1st percentile lower confidence limit</td>
</tr>
<tr>
<td>P1_UCL</td>
<td>1st percentile upper confidence limit</td>
</tr>
<tr>
<td>P5_LCL</td>
<td>5th percentile lower confidence limit</td>
</tr>
<tr>
<td>P5_UCL</td>
<td>5th percentile upper confidence limit</td>
</tr>
<tr>
<td>P10_LCL</td>
<td>10th percentile lower confidence limit</td>
</tr>
<tr>
<td>P10_UCL</td>
<td>10th percentile upper confidence limit</td>
</tr>
<tr>
<td>Q1_LCL</td>
<td></td>
</tr>
<tr>
<td>Q2_LCL</td>
<td>Lower quartile (25th percentile) lower confidence limit</td>
</tr>
<tr>
<td>Q1_UCL</td>
<td>Lower quartile (25th percentile) upper confidence limit</td>
</tr>
<tr>
<td>MEDIAN_LCL</td>
<td>Lower quartile (25th percentile) upper confidence limit</td>
</tr>
<tr>
<td>Q2_UCL</td>
<td></td>
</tr>
<tr>
<td>P50_LCL</td>
<td>Median (50th percentile) lower confidence limit</td>
</tr>
<tr>
<td>MEDIAN_UCL</td>
<td>Median (50th percentile) upper confidence limit</td>
</tr>
<tr>
<td>Q2_UCL</td>
<td></td>
</tr>
<tr>
<td>P50_UCL</td>
<td>Median (50th percentile) upper confidence limit</td>
</tr>
<tr>
<td>Q3_LCL</td>
<td></td>
</tr>
<tr>
<td>P75_LCL</td>
<td>Upper quartile (75th percentile) lower confidence limit</td>
</tr>
<tr>
<td>Q3_UCL</td>
<td>Upper quartile (75th percentile) upper confidence limit</td>
</tr>
<tr>
<td>P75_UCL</td>
<td>Upper quartile (75th percentile) upper confidence limit</td>
</tr>
<tr>
<td>P90_LCL</td>
<td>90th percentile lower confidence limit</td>
</tr>
<tr>
<td>P90_UCL</td>
<td>90th percentile upper confidence limit</td>
</tr>
</tbody>
</table>
Table 4.10 continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P95_LCL</td>
<td>95th percentile lower confidence limit</td>
</tr>
<tr>
<td>P95_UCL</td>
<td>95th percentile upper confidence limit</td>
</tr>
<tr>
<td>P99_LCL</td>
<td>99th percentile lower confidence limit</td>
</tr>
<tr>
<td>P99_UCL</td>
<td>99th percentile upper confidence limit</td>
</tr>
</tbody>
</table>

Robust Statistics Keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GINI</td>
<td>Gini’s mean difference</td>
</tr>
<tr>
<td>MAD</td>
<td>Median absolute difference about the median</td>
</tr>
<tr>
<td>QN</td>
<td>$Q_n$, alternative to MAD</td>
</tr>
<tr>
<td>SN</td>
<td>$S_n$, alternative to MAD</td>
</tr>
<tr>
<td>STD_GINI</td>
<td>Gini’s standard deviation</td>
</tr>
<tr>
<td>STD_MAD</td>
<td>MAD standard deviation</td>
</tr>
<tr>
<td>STD_QN</td>
<td>$Q_n$ standard deviation</td>
</tr>
<tr>
<td>STD_QRANGE</td>
<td>Interquartile range standard deviation</td>
</tr>
<tr>
<td>STD_SN</td>
<td>$S_n$ standard deviation</td>
</tr>
</tbody>
</table>

Hypothesis Testing Keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSIGN</td>
<td>Sign statistic</td>
</tr>
<tr>
<td>NORMALTEST</td>
<td>Test statistic for normality</td>
</tr>
<tr>
<td>PNORMAL</td>
<td>Probability value for the test of normality</td>
</tr>
<tr>
<td>SIGNRANK</td>
<td>Signed rank statistic</td>
</tr>
<tr>
<td>PROBM</td>
<td>Probability of greater absolute value for the sign statistic</td>
</tr>
<tr>
<td>PROBN</td>
<td>Probability value for the test of normality</td>
</tr>
<tr>
<td>PROBS</td>
<td>Probability value for the signed rank test</td>
</tr>
<tr>
<td>PROBT</td>
<td>Probability value for the Student’s $t$ test</td>
</tr>
<tr>
<td>T</td>
<td>Statistics for Student’s $t$ test</td>
</tr>
</tbody>
</table>

Keyword for Reading an Input Data Set

DATA= (label, value) pairs from input data set

To create a completely customized inset, use a DATA= data set.

**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_ is a character variable whose values provide labels for inset entries.
- _VALUE_ is 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.
Primary and Secondary Keywords

A primary keyword specifies a fitted distribution, which is one of the parametric distributions or a kernel density estimate. You specify secondary keywords in parentheses after the primary keyword to request particular statistics associated with that distribution.

**NOTE:** When producing traditional graphics output, you can specify a primary keyword without secondary keywords to display a colored line and the distribution name as a key for the density curve.

In the HISTOGRAM statement you can request more than one fitted distribution from the same family (for example, two normal distributions). You can display inset statistics for individual curves by specifying the curve indices in square brackets immediately following the primary keyword.

The following statements produce a histogram with three fitted normal curves and an inset that contains goodness-of-fit statistics for the second curve only:

```plaintext
proc univariate data=score;
    histogram final / normal(sigma=1 2 3);
    inset normal[2](ad adpval);
run;
```

Table 4.11 lists the primary keywords and the plot statements with which they can be specified.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Distribution</th>
<th>Plot Statement Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA</td>
<td>Beta</td>
<td>All plot statements</td>
</tr>
<tr>
<td>EXPONENTIAL</td>
<td>Exponential</td>
<td>All plot statements</td>
</tr>
<tr>
<td>GAMMA</td>
<td>Gamma</td>
<td>All plot statements</td>
</tr>
<tr>
<td>GUMBEL</td>
<td>Gumbel</td>
<td>All plot statements</td>
</tr>
<tr>
<td>IGAUSS</td>
<td>Inverse Gaussian</td>
<td>CDFPLOT, HISTOGRAM, PPLOT</td>
</tr>
<tr>
<td>KERNEL</td>
<td>Kernel density estimate</td>
<td>HISTOGRAM</td>
</tr>
<tr>
<td>LOGNORMAL</td>
<td>Lognormal</td>
<td>All plot statements</td>
</tr>
<tr>
<td>NORMAL</td>
<td>Normal</td>
<td>All plot statements</td>
</tr>
<tr>
<td>PARETO</td>
<td>Pareto</td>
<td>All plot statements</td>
</tr>
<tr>
<td>POWER</td>
<td>Power function</td>
<td>All plot statements</td>
</tr>
<tr>
<td>RAYLEIGH</td>
<td>Rayleigh</td>
<td>All plot statements</td>
</tr>
<tr>
<td>SB</td>
<td>Johnson $S_B$</td>
<td>HISTOGRAM</td>
</tr>
<tr>
<td>SU</td>
<td>Johnson $S_U$</td>
<td>HISTOGRAM</td>
</tr>
<tr>
<td>WEIBULL</td>
<td>Weibull(3-parameter)</td>
<td>All plot statements</td>
</tr>
<tr>
<td>WEIBULL2</td>
<td>Weibull(2-parameter)</td>
<td>PROBPLOT, QQPLOT</td>
</tr>
</tbody>
</table>

Table 4.12 lists the secondary keywords available with the primary keywords listed in Table 4.11.

<table>
<thead>
<tr>
<th>Secondary Keyword</th>
<th>Alias</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA Secondary Keywords</td>
<td></td>
<td>First shape parameter $\alpha$</td>
</tr>
<tr>
<td>ALPHA</td>
<td>SHAPE1</td>
<td></td>
</tr>
<tr>
<td>BETA</td>
<td>SHAPE2</td>
<td>Second shape parameter $\beta$</td>
</tr>
</tbody>
</table>
### Table 4.12  continued

<table>
<thead>
<tr>
<th>Secondary Keyword</th>
<th>Alias</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Lower threshold parameter $\theta$</td>
</tr>
</tbody>
</table>

**EXPONENTIAL Secondary Keywords**
- **MEAN**: Mean of the fitted distribution
- **SIGMA**: Scale parameter $\sigma$
- **STD**: Standard deviation of the fitted distribution
- **THETA**: Threshold parameter $\theta$

**GAMMA Secondary Keywords**
- **ALPHA**: Shape parameter $\alpha$
- **MEAN**: Mean of the fitted distribution
- **SIGMA**: Scale parameter $\sigma$
- **STD**: Standard deviation of the fitted distribution
- **THETA**: Threshold parameter $\theta$

**GUMBEL Secondary Keywords**
- **MEAN**: Mean of the fitted distribution
- **MU**: Location parameter $\mu$
- **SIGMA**: Scale parameter $\sigma$
- **STD**: Standard deviation of the fitted distribution

**IGAUSS Secondary Keywords**
- **LAMBDA**: Shape parameter $\lambda$
- **MEAN**: Mean of the fitted distribution
- **MU**: Mean parameter $\mu$
- **STD**: Standard deviation of the fitted distribution

**KERNEL Secondary Keywords**
- **AMISE**: Approximate mean integrated square error (MISE) for the kernel density
- **BANDWIDTH**: Bandwidth $\lambda$ for the density estimate
- **BWIDTH**: Alias for BANDWIDTH
- **C**: Standardized bandwidth for the density estimate
- **TYPE**: Kernel type: normal, quadratic, or triangular

**LOGNORMAL Secondary Keywords**
- **MEAN**: Mean of the fitted distribution
- **SIGMA**: Shape parameter $\sigma$
- **STD**: Standard deviation of the fitted distribution
- **THETA**: Threshold parameter $\theta$
- **ZETA**: Scale parameter $\zeta$
### Table 4.12  continued

<table>
<thead>
<tr>
<th>Secondary Keyword</th>
<th>Alias</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NORMAL Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MU</td>
<td>MEAN</td>
<td>Mean parameter $\mu$</td>
</tr>
<tr>
<td>SIGMA</td>
<td>STD</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td><strong>PARETO Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALPHA</td>
<td></td>
<td>Shape parameter $\alpha$</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>POWER Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALPHA</td>
<td></td>
<td>Shape parameter $\alpha$</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>RAYLEIGH Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>SB and SU Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DELTA</td>
<td>SHAPE1</td>
<td>First shape parameter $\delta$</td>
</tr>
<tr>
<td>GAMMA</td>
<td>SHAPE2</td>
<td>Second shape parameter $\gamma$</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Lower threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>WEIBULL Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>SHAPE</td>
<td>Shape parameter $c$</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>WEIBULL2 Secondary Keywords</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>SHAPE</td>
<td>Shape parameter $c$</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td>Mean of the fitted distribution</td>
</tr>
<tr>
<td>SIGMA</td>
<td>SCALE</td>
<td>Scale parameter $\sigma$</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td>Standard deviation of the fitted distribution</td>
</tr>
<tr>
<td>THETA</td>
<td>THRESHOLD</td>
<td>Known lower threshold $\theta_0$</td>
</tr>
</tbody>
</table>
The inset statistics listed in Table 4.12 are not available unless you request a plot statement and options that calculate these statistics. For example, consider the following statements:

```plaintext
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, respectively, of `final`. The NORMAL keyword with the secondary keywords AD and ADPV AL displays the Anderson-Darling goodness-of-fit test statistic and $p$-value, respectively. The statistics that are specified with the NORMAL keyword are available only because the NORMAL option is requested in the HISTOGRAM statement.

The KERNEL 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.

### INSET Statistic Labels and Formats

By default, PROC UNIVARIATE identifies inset statistics with appropriate labels and prints numeric values with 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, the following statement requests customized labels for two statistics and displays the standard deviation with a field width of 5 and two decimal places:

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

### Summary of Options

Table 4.13 lists INSET statement `options`, which are specified after the slash (/) in the INSET statement. For complete descriptions, see the section “Dictionary of Options” on page 324.

---

**Table 4.12 continued**

<table>
<thead>
<tr>
<th>Secondary Keyword</th>
<th>Alias</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td></td>
<td>Anderson-Darling EDF test statistic</td>
</tr>
<tr>
<td>ADPV AL</td>
<td></td>
<td>Anderson-Darling EDF test $p$-value</td>
</tr>
<tr>
<td>CVM</td>
<td></td>
<td>Cramér–von Mises EDF test statistic</td>
</tr>
<tr>
<td>CVMPV AL</td>
<td></td>
<td>Cramér–von Mises EDF test $p$-value</td>
</tr>
<tr>
<td>KSD</td>
<td></td>
<td>Kolmogorov-Smirnov EDF test statistic</td>
</tr>
<tr>
<td>KSDPV AL</td>
<td></td>
<td>Kolmogorov-Smirnov EDF test $p$-value</td>
</tr>
</tbody>
</table>
Table 4.13  INSET Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFILL=color \ BLANK</td>
<td>specifies color of inset background</td>
</tr>
<tr>
<td>CFILLH=color</td>
<td>specifies color of header background</td>
</tr>
<tr>
<td>CFRAME=color</td>
<td>specifies color of frame</td>
</tr>
<tr>
<td>CHEADER=color</td>
<td>specifies color of header text</td>
</tr>
<tr>
<td>CSHADOW=color</td>
<td>specifies color of drop shadow</td>
</tr>
<tr>
<td>CTEXT=color</td>
<td>specifies color of inset text</td>
</tr>
<tr>
<td>DATA</td>
<td>specifies data units for POSITION=(x, y) coordinates</td>
</tr>
<tr>
<td>FONT=font</td>
<td>specifies font of text</td>
</tr>
<tr>
<td>FORMAT=format</td>
<td>specifies format of values in inset</td>
</tr>
<tr>
<td>GUTTER=value</td>
<td>specifies gutter width for inset in top or bottom margin</td>
</tr>
<tr>
<td>HEADER='string'</td>
<td>specifies header text</td>
</tr>
<tr>
<td>HEIGHT=value</td>
<td>specifies height of inset text</td>
</tr>
<tr>
<td>NCOLS=</td>
<td>specifies number of columns for inset in top or bottom margin</td>
</tr>
<tr>
<td>NOFRAME</td>
<td>suppresses frame around inset</td>
</tr>
<tr>
<td>POSITION=position</td>
<td>specifies position of inset</td>
</tr>
<tr>
<td>REFPOINT=BR</td>
<td>BL</td>
</tr>
</tbody>
</table>

Dictionary of Options

The following entries provide detailed descriptions of options for the INSET statement. Options marked with † are applicable only when traditional graphics are produced.

† CFILL=color \ BLANK

specifies the color of the background for traditional graphics. If you omit the CFILLH= option the header background is included. By default, the background is empty, which causes items that overlap the inset (such as curves or histogram bars) to show through the inset.

If you specify a value for CFILL= option, then overlapping items no longer show through the inset. Use CFILL=BLANK to leave the background uncolored and to prevent items from showing through the inset.

† CFILLH=color

specifies the color of the header background for traditional graphics. The default value is the CFILL= color.

† CFRAME=color

specifies the color of the frame for traditional graphics. The default value is the same color as the axis of the plot.

† CHEADER=color

specifies the color of the header text for traditional graphics. The default value is the CTEXT= color.

† CSHADOW=color

specifies the color of the drop shadow for traditional graphics. By default, if a CSHADOW= option is not specified, a drop shadow is not displayed.
† CTEXT=color
specifies the color of the text for traditional graphics. The default value is the same color as the other text on the plot.

DATA
specifies that data coordinates are to be used in positioning the inset with the POSITION= option. The DATA option is available only when you specify POSITION=(x,y). You must place DATA immediately after the coordinates (x,y). **NOTE:** Positioning insets with coordinates is not supported for ODS Graphics output.

† FONT=font
specifies the font of the text for traditional graphics. By 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.

FORMAT=format
specifies a format for all the values in the inset. If you specify a format for a particular statistic, then that format overrides the one specified with the FORMAT= option. For more information about SAS formats, see *SAS Formats and Informats: Reference*.

GUTTER=value
specifies the gutter width in percent screen units for an inset located in the top or bottom margin. The gutter is the space between columns of (label, value) pairs in an inset. The default value is four. **NOTE:** The GUTTER= option applies only when ODS Graphics is enabled.

HEADER=string
specifies the header text. The string cannot exceed 40 characters. By default, no header line appears in the inset. 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.

† HEIGHT=value
specifies the height of the text for traditional graphics.

NCOLS=n
specifies the number of columns of (label, value) pairs displayed in an inset located in the top or bottom margin. The default value is three. **NOTE:** The NCOLS= option applies only when ODS Graphics is enabled.

NOFRAME
suppresses the frame drawn around the text.

POSITION=position
POS=position
determines the position of the inset. The position is a compass point keyword, a margin keyword, or a pair of coordinates (x,y). You can specify coordinates in axis percent units or axis data units. The default value is NW, which positions the inset in the upper left (northwest) corner of the display. See the section “Positioning Insets” on page 397.

**NOTE:** Positioning insets with coordinates is not supported for ODS Graphics output.
† 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 keywords are BL, BR, TL, and TR, which correspond to bottom left, bottom right, top left, and top right. The default value is BL. You must use REFPOINT= with POSITION=(x,y) coordinates. The option does not apply to ODS Graphics output.

---

**OUTPUT Statement**

```
OUTPUT < OUT=SAS-data-set > < keyword1=names . . . keywordk=names > < percentile-options > ;
```

The OUTPUT statement saves statistics and BY variables in an output data set. When you use a BY statement, each observation in the output data set corresponds to one of the BY groups. Otherwise, the output data set contains only one observation.

You can use any number of OUTPUT statements in the UNIVARIATE procedure. Each OUTPUT statement creates a new data set to contain the statistics specified in that statement. You must use the VAR statement with the OUTPUT statement. The OUTPUT statement must contain a specification of the form `keyword=names` or the PCTLPTS= and PCTLPRE= options. See Example 4.7 and Example 4.8.

You can use the OUT= option to specify the name of the output data set:

```
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.

A `keyword=names` specification selects a statistic to be included in the output data set and gives the names of new variables that contain the statistic. Specify a `keyword` for each desired statistic, followed by an equal sign, followed by the `names` of the variables to contain the statistic. In the output data set, the first variable listed after a keyword in the OUTPUT statement contains the statistic for the first variable listed in the VAR statement, the second variable contains the statistic for the second variable in the VAR statement, and so on. If the list of `names` following the equal sign is shorter than the list of variables in the VAR statement, the procedure uses the `names` in the order in which the variables are listed in the VAR statement. The available keywords are listed in Table 4.14.

**Table 4.14** Statistical Keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive Statistic Keywords</td>
<td></td>
</tr>
<tr>
<td>CSS</td>
<td>Corrected sum of squares</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of variation</td>
</tr>
<tr>
<td>KURTOSIS</td>
<td>KURT</td>
</tr>
<tr>
<td>MAX</td>
<td>Largest value</td>
</tr>
<tr>
<td>MEAN</td>
<td>Sample mean</td>
</tr>
<tr>
<td>MIN</td>
<td>Smallest value</td>
</tr>
<tr>
<td>MODE</td>
<td>Most frequent value</td>
</tr>
<tr>
<td>N</td>
<td>Sample size</td>
</tr>
<tr>
<td>NMISS</td>
<td>Number of missing values</td>
</tr>
</tbody>
</table>

---
Table 4.14  continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOBS</td>
<td>Number of observations</td>
</tr>
<tr>
<td>RANGE</td>
<td>Range</td>
</tr>
<tr>
<td>SKEWNESS</td>
<td>SKEW</td>
</tr>
<tr>
<td>STD</td>
<td>STDDEV</td>
</tr>
<tr>
<td>STDMEAN</td>
<td>STDERR</td>
</tr>
<tr>
<td>SUM</td>
<td>Sum of the observations</td>
</tr>
<tr>
<td>SUMWGT</td>
<td>Sum of the weights</td>
</tr>
<tr>
<td>USS</td>
<td>Uncorrected sum of squares</td>
</tr>
<tr>
<td>VAR</td>
<td>Variance</td>
</tr>
</tbody>
</table>

**Quantile Statistic Keywords**

| P1               | 1st percentile                       |
| P5               | 5th percentile                       |
| P10              | 10th percentile                      |
| Q1 | P25           | Lower quartile (25th percentile)     |
| MEDIAN | Q2 | P50         | Median (50th percentile)            |
| Q3 | P75           | Upper quartile (75th percentile)     |
| P90              | 90th percentile                      |
| P95              | 95th percentile                      |
| P99              | 99th percentile                      |
| QRANGE           | Interquartile range (Q3–Q1)          |

**Robust Statistic Keywords**

| GINI            | Gini’s mean difference               |
| MAD             | Median absolute difference about the median |
| QN             | $Q_n$, alternative to MAD            |
| SN             | $S_n$, alternative to MAD            |
| STD_GINI       | Gini’s standard deviation            |
| STD_MAD        | MAD standard deviation                |
| STD_QN         | $Q_n$ standard deviation             |
| STD_QRANGE     | Interquartile range standard deviation|
| STD_SN         | $S_n$ standard deviation              |

**Hypothesis Testing Keywords**

| MSIGN          | Sign statistic                       |
| NORMALTEST     | Test statistic for normality         |
| SIGNRANK       | Signed rank statistic                |
| PROBM          | Probability of a greater absolute value for the sign statistic |
| PROBN          | Probability value for the test of normality |
| PROBS          | Probability value for the signed rank test |
| PROBT          | Probability value for the Student’s $t$ test |
| T              | Statistic for the Student’s $t$ test  |
The UNIVARIATE procedure automatically computes the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles for the data. These can be saved in an output data set by using `keyword=names` specifications. You can request additional percentiles by using the `PCTLPTS=` option. The following percentile-options are related to these additional percentiles:

CIPCTLDF=(cipctl-options)
CIQUANTDF=(cipctl-options)

requests distribution-free confidence limits for percentiles that are requested with the `PCTLPTS=` option. 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). This option does not apply if you use a WEIGHT statement. You can specify the following cipctl-options:

\[ \text{ALPHA=} \alpha \]

specifies the level of significance \( \alpha \) for \( 100(1 - \alpha)\% \) confidence intervals. The value \( \alpha \) must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

\[ \text{LOWERPRE=} \text{prefixes} \]

specifies one or more prefixes that are used to create names for variables that contain the lower confidence limits. To save lower confidence limits 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.

\[ \text{LOWERNAME=} \text{suffixes} \]

specifies one or more suffixes that are used to create names for variables that contain the lower confidence limits. PROC UNIVARIATE creates a variable name by combining the LOWERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the `PCTLPTS=` percentiles.

\[ \text{TYPE=} \text{keyword} \]

specifies the type of confidence limit, where `keyword` is LOWER, UPPER, SYMMETRIC, or ASYMMETRIC. The default value is SYMMETRIC.

\[ \text{UPPERPRE=} \text{prefixes} \]

specifies one or more prefixes that are used to create names for variables that contain the upper confidence limits. To save upper confidence limits 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.

\[ \text{UPPERNAME=} \text{suffixes} \]

specifies one or more suffixes that are used to create names for variables that contain the upper confidence limits. PROC UNIVARIATE creates a variable name by combining the UPPERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the `PCTLPTS=` percentiles.

**Note:** See the entries for the `PCTLPTS=`, `PCTLPRE=`, and `PCTLNAME=` options for a detailed description of how variable names are created using prefixes, percentile values, and suffixes.
requests confidence limits based on the assumption that the data are normally distributed for percentiles that are requested with the PCTLPTS= option. The computational method is described in Section 4.4.1 of Hahn and Meeker (1991) and uses the noncentral t distribution as given by Odeh and Owen (1980). This option does not apply if you use a WEIGHT statement. You can specify the following cipctl-options:

**ALPHA=α**

specifies the level of significance $\alpha$ for $100(1 - \alpha)\%$ confidence intervals. The value $\alpha$ must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

**LOWERPRE=prefixes**

 specifies one or more prefixes that are used to create names for variables that contain the lower confidence limits. To save lower confidence limits 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.

**LOWERNAME=suffixes**

 specifies one or more suffixes that are used to create names for variables that contain the lower confidence limits. PROC UNIVARIATE creates a variable name by combining the LOWERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the PCTLPTS= percentiles.

**TYPE=keyword**

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

**UPPERPRE=prefixes**

 specifies one or more prefixes that are used to create names for variables that contain the upper confidence limits. To save upper confidence limits 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.

**UPPERNAME=suffixes**

 specifies one or more suffixes that are used to create names for variables that contain the upper confidence limits. PROC UNIVARIATE creates a variable name by combining the UPPERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the PCTLPTS= percentiles.

**NOTE:** See the entries for the PCTLPTS=, PCTLPRE=, and PCTLNAME= options for a detailed description of how variable names are created using prefixes, percentile values, and suffixes.

**PCTLGROUP=BYSTAT | BYVAR**

specifies the order in which variables that you request with the PCTLPTS= option are added to the OUT= data set when the VAR statement lists more than one analysis variable. By default (or if you specify PCTLGROUP=BYSTAT), all variables that are associated with a percentile value are created consecutively. If you specify PCTLGROUP=BYVAR, all variables that are associated with an analysis variable are created consecutively.

Consider the following statements:
proc univariate data=Score;
  var PreTest PostTest;
  output out=ByStat pctlpts=20 40 pctlpre=Pre_Post;
  output out=ByVar pctlgroup=byvar pctlpts=20 40 pctlpre=Pre_Post;
run;

The order of variables in the data set ByStat is Pre_20, Post_20, Pre_40, Post_40. The order of variables in the data set ByVar is Pre_20, Pre_40, Post_20, Post_40.

**PCTLNAME=suffixes**

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. If you specify \( n \) suffixes with the PCTLNAME= option and \( m \) percentile values with the PCTLPTS= option where \( m > n \), the suffixes are used to name the first \( n \) percentiles and the default names are used for the remaining \( m - n \) percentiles. For example, consider the following statements:

```
proc univariate;
  var Length Width Height;
  output pctlpts = 20 40
    pctlpre = pl pw ph
    pctlname = twenty;
run;
```

The value *twenty* in the PCTLNAME= option is used for only the first percentile in the PCTLPTS= list. This suffix is appended to the values in the PCTLPRE= option to generate the new variable names pltwenty, pwtwenty, and phtwenty, which contain the 20th percentiles for Length, Width, and Height, respectively. Because a second PCTLNAME= suffix is not specified, variable names for the 40th percentiles for Length, Width, and Height are generated using the prefixes and percentile values. Thus, the output data set contains the variables pltwenty, pl40, pwtwenty, pw40, phtwenty, and ph40.

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

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

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.

**PCTLNDEC=value**

specifies the number of decimal places in percentile values that are incorporated into percentile variable names. The default value is 2. For example, the following statements create two output data sets, each containing one percentile variable. The variable in data set short is named pwid85_12, while the one in data set long is named pwid85_125.
proc univariate;
  var width;
  output out=short pctlpts=85.125 pctlpre=pwid;
  output out=long pctlpts=85.125 pctlpre=pwid pctlndec=3;
run;

PCTLPRE=prefixes
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. The PCTLPRE= and PCTLPTS= options must be used together.

The procedure generates new variable names by using the prefix and the percentile values. If the specified percentile is an integer, the variable name is simply the prefix followed by the value. If the specified value is not an integer, an underscore replaces the decimal point in the variable name, and decimal values are truncated to one decimal place. For example, the following statements create the variables pwid20, pwid33.3, pwid66.7, and pwid80 for the 20th, 33.3rd, 66.7th, and 80th percentiles of Width, respectively:

proc univariate noprint;
  var Width;
  output pctlpts=20 33.33 66.67 80 pctlpre=pwid;
run;

If you request percentiles for more than one variable, you should list prefixes in the same order in which the variables appear in the VAR statement. If combining the prefix and percentile value results in a name longer than 32 characters, the prefix is truncated so that the variable name is 32 characters.

PCTLPTS=percentiles
specifies one or more percentiles that are not automatically computed by the UNIVARIATE procedure. The PCTLPRE= and PCTLPTS= options must be used together. You can specify percentiles with an expression of the form start TO stop BY increment where start is a starting number, stop is an ending number, and increment is a number to increment by. The PCTLPTS= option generates additional percentiles and outputs them to a data set. These additional percentiles are not printed.

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

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

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 named Pctls that contains the 20th and 40th percentiles of the analysis variables PreTest and PostTest:

proc univariate data=Score;
  var PreTest PostTest;
  output out=Pctls pctlpts=20 40 pctlpre=PreTest_ PostTest_;
pctlname=P20 P40;
run;

PROC UNIVARIA TE saves the 20th and 40th percentiles for PreTest and PostTest in the variables PreTest_P20, PostTest_P20, PreTest_P40, and PostTest_P40.

### PPPLOT Statement

**PPPLOT < variables > < / options > ;**

The PPPLOT statement creates a probability-probability plot (also referred to as a P-P plot or percent plot), which compares the empirical cumulative distribution function (ECDF) of a variable with a specified theoretical cumulative distribution function such as the normal. If the two distributions match, the points on the plot form a linear pattern that passes through the origin and has unit slope. Thus, you can use a P-P plot to determine how well a theoretical distribution models a set of measurements.

You can specify one of the following theoretical distributions with the PPPLOT statement:

- beta
- exponential
- gamma
- Gumbel
- generalized Pareto
- inverse Gaussian
- lognormal
- normal
- power function
- Rayleigh
- Weibull

**NOTE:** Probability-probability plots should not be confused with probability plots, which compare a set of ordered measurements with percentiles from a specified distribution. You can create probability plots with the PROBPLOT statement.

You can use any number of PPPLOT statements in the UNIVARIATE procedure. The components of the PPPLOT statement are as follows.

**variables**

are the process variables for which P-P plots are created. If you specify a VAR statement, the **variables** must also be listed in the VAR statement. Otherwise, the **variables** can be any numeric variables in the input data set. If you do not specify a list of **variables**, then by default, the procedure creates a P-P plot for each variable listed in the VAR statement or for each numeric variable in the input data set if you do not specify a VAR statement. For example, if data set **measures** contains two numeric variables, **length** and **width**, the following two PPPLOT statements each produce a P-P plot for each of those variables:
options

specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all options after the slash (/) in the PPLOT statement. You can specify only one option that names a distribution, but you can specify any number of other options. By default, the procedure produces a P-P plot based on the normal distribution.

In the following example, the NORMAL, MU=, and SIGMA= options request a P-P plot based on the normal distribution with mean 10 and standard deviation 0.3. The SQUARE option displays the plot in a square frame, and the CTEXT= option specifies the text color.

Table 4.15 through Table 4.17 list the PPLOT options by function. For complete descriptions, see the sections “Dictionary of Options” on page 337 and “Dictionary of Common Options” on page 369. Options can be any of the following:

- primary options
- secondary options
- general options

Distribution Options

Table 4.15 summarizes the options for requesting a specific theoretical distribution.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA(beta-options)</td>
<td>specifies beta P-P plot</td>
</tr>
<tr>
<td>EXPONENTIAL(exponential-options)</td>
<td>specifies exponential P-P plot</td>
</tr>
<tr>
<td>GAMMA(gamma-options)</td>
<td>specifies gamma P-P plot</td>
</tr>
<tr>
<td>GUMBEL(Gumbel-options)</td>
<td>specifies Gumbel P-P plot</td>
</tr>
<tr>
<td>PARETO(Pareto-options)</td>
<td>specifies generalized Pareto P-P plot</td>
</tr>
<tr>
<td>IGAUSS(iGauss-options)</td>
<td>specifies inverse Gaussian P-P plot</td>
</tr>
<tr>
<td>LOGNORMAL(lognormal-options)</td>
<td>specifies lognormal P-P plot</td>
</tr>
</tbody>
</table>
Table 4.15 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMAL(normal-options)</td>
<td>specifies normal P-P plot</td>
</tr>
<tr>
<td>POWER(power-options)</td>
<td>specifies power function P-P plot</td>
</tr>
<tr>
<td>RAYLEIGH(Rayleigh-options)</td>
<td>specifies Rayleigh P-P plot</td>
</tr>
<tr>
<td>WEIBULL(Weibull-options)</td>
<td>specifies Weibull P-P plot</td>
</tr>
</tbody>
</table>

Table 4.16 summarizes options that specify distribution parameters and control the display of the diagonal distribution reference line. Specify these options in parentheses after the distribution option. For example, the following statements use the NORMAL option to request a normal P-P plot:

```plaintext
proc univariate data=measures;
   ppplot length / normal(mu=10 sigma=0.3 color=red);
run;
```

The MU= and SIGMA= normal-options specify $\mu$ and $\sigma$ for the normal distribution, and the COLOR= normal-option specifies the color for the line.

Table 4.16 Secondary Distribution Reference Line Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR=</td>
<td>specifies color of distribution reference line</td>
</tr>
<tr>
<td>L=</td>
<td>specifies line type of distribution reference line</td>
</tr>
<tr>
<td>NOLINE</td>
<td>suppresses the distribution reference line</td>
</tr>
<tr>
<td>W=</td>
<td>specifies width of distribution reference line</td>
</tr>
<tr>
<td><strong>Beta-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$</td>
</tr>
<tr>
<td>BETA=</td>
<td>specifies shape parameter $\beta$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies lower threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>Exponential-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>Gamma-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$</td>
</tr>
<tr>
<td>ALPHADELTA=</td>
<td>specifies change in successive estimates of $\alpha$ at which the Newton-Raphson approximation of $\hat{\alpha}$ terminates</td>
</tr>
<tr>
<td>ALPHAINITIAL=</td>
<td>specifies initial value for $\alpha$ in the Newton-Raphson approximation of $\hat{\alpha}$</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{\alpha}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
<tr>
<td><strong>Gumbel-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies location parameter $\mu$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td><strong>IGauss-Options</strong></td>
<td></td>
</tr>
<tr>
<td>LAMBDA=</td>
<td>specifies shape parameter $\lambda$</td>
</tr>
<tr>
<td>MU=</td>
<td>specifies mean $\mu$</td>
</tr>
</tbody>
</table>
### Lognormal-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGMA=</td>
<td>specifies shape parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
<tr>
<td>ZETA=</td>
<td>specifies scale parameter $\zeta$</td>
</tr>
</tbody>
</table>

### Normal-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MU=</td>
<td>specifies mean $\mu$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies standard deviation $\sigma$</td>
</tr>
</tbody>
</table>

### Pareto-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
</tbody>
</table>

### Power-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPHA=</td>
<td>specifies shape parameter $\alpha$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
</tbody>
</table>

### Rayleigh-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
</tbody>
</table>

### Weibull-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=</td>
<td>specifies shape parameter $c$</td>
</tr>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{c}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies scale parameter $\sigma$</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies threshold parameter $\theta$</td>
</tr>
</tbody>
</table>

---

### General Options

Table 4.17 lists options that control the appearance of the plots. For complete descriptions, see the sections “Dictionary of Options” on page 337 and “Dictionary of Common Options” on page 369.

#### Table 4.17 General Graphics Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HREF=</td>
<td>specifies reference lines perpendicular to the horizontal axis</td>
</tr>
<tr>
<td>HREFLABELS=</td>
<td>specifies line labels for HREF= lines</td>
</tr>
<tr>
<td>HREFLABPOS=</td>
<td>specifies position for HREF= line labels</td>
</tr>
<tr>
<td>NOHLABEL</td>
<td>suppresses label for horizontal axis</td>
</tr>
<tr>
<td>NOVLABEL</td>
<td>suppresses label for vertical axis</td>
</tr>
<tr>
<td>NOVTICK</td>
<td>suppresses tick marks and tick mark labels for vertical axis</td>
</tr>
<tr>
<td>SQUARE</td>
<td>displays P-P plot in square format</td>
</tr>
<tr>
<td>VAXISLABEL=</td>
<td>specifies label for vertical axis</td>
</tr>
<tr>
<td>VREF=</td>
<td>specifies reference lines perpendicular to the vertical axis</td>
</tr>
<tr>
<td>Option</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>VREFLABELS=</td>
<td>specifies line labels for VREF= lines</td>
</tr>
<tr>
<td>VREFLABPOS=</td>
<td>specifies position for VREF= line labels</td>
</tr>
</tbody>
</table>

**Options for Traditional Graphics Output**
- ANNOTATE= provides an annotate data set
- CAXIS= specifies color for axis
- CFRAME= specifies color for frame
- CHREF= specifies colors for HREF= lines
- CTEXT= specifies color for text
- CVREF= specifies colors for VREF= lines
- DESCRIPTION= specifies description for plot in graphics catalog
- FONT= specifies software font for text
- HAXIS= specifies AXIS statement for horizontal axis
- HEIGHT= specifies height of text used outside framed areas
- HMINOR= specifies number of minor tick marks on horizontal axis
- INFONT= specifies software font for text inside framed areas
- INHEIGHT= specifies height of text inside framed areas
- LHREF= specifies line types for HREF= lines
- LVREF= specifies line types for VREF= lines
- NAME= specifies name for plot in graphics catalog
- NOFRAME suppresses frame around plotting area
- TURNVLABELS turns and vertically strings out characters in labels for vertical axis
- VAXIS= specifies AXIS statement for vertical axis
- VMINOR= specifies number of minor tick marks on vertical axis
- WAXIS= specifies line thickness for axes and frame

**Options for ODS Graphics Output**
- ODSFOOTNOTE= specifies footnote displayed on plot
- ODSFOOTNOTE2= specifies secondary footnote displayed on plot
- ODSTITLE= specifies title displayed on plot
- ODSTITLE2= specifies secondary title displayed on plot
- OVERLAY overlays plots for different class levels (ODS Graphics only)

**Options for Comparative Plots**
- ANNOKEY applies annotation requested in ANNOTATE= data set to key cell only
- CFRAMESIDE= specifies color for filling row label frames
- CFRAMETOP= specifies color for filling column label frames
- CPROP= specifies color for proportion of frequency bar
- CTEXTSIDE= specifies color for row labels
- CTEXTTOP= specifies color for column labels
- INTERTILE= specifies distance between tiles in comparative plot
- NCOLS= specifies number of columns in comparative plot
- NROWS= specifies number of rows in comparative plot

**Miscellaneous Options**
- CONTENTS= specifies table of contents entry for P-P plot grouping
Dictionary of Options

The following entries provide detailed descriptions of options for the PPLOT statement. See the section “Dictionary of Common Options” on page 369 for detailed descriptions of options common to all plot statements.

**ALPHA=value**

specifies the shape parameter \( \alpha \) \((\alpha > 0)\) for P-P plots requested with the BETA, GAMMA, PARETO and POWER options.

**BETA< (beta-options) >**

creates a beta P-P plot. To create the plot, the \( n \) nonmissing observations are ordered from smallest to largest:

\[
x(1) \leq x(2) \leq \cdots \leq x(n)
\]

The \( y \)-coordinate of the \( i \)th point is the empirical cdf value \( \frac{i}{n} \). The \( x \)-coordinate is the theoretical beta cdf value

\[
B_{\alpha\beta}\left(\frac{x(i) - \theta}{\sigma}\right) = \int_{\theta}^{x(i)} \frac{(t - \theta)^{\alpha-1}(\theta + \sigma - t)^{\beta-1}}{B(\alpha, \beta)\sigma^{\alpha+\beta-1}} dt
\]

where \( B_{\alpha\beta}(\cdot) \) is the normalized incomplete beta function, \( B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \), and

- \( \theta = \) lower threshold parameter
- \( \sigma = \) scale parameter \((\sigma > 0)\)
- \( \alpha = \) first shape parameter \((\alpha > 0)\)
- \( \beta = \) second shape parameter \((\beta > 0)\)

You can specify \( \alpha, \beta, \sigma, \) and \( \theta \) with the ALPHA=, BETA=, SIGMA=, and THETA= beta-options, as illustrated in the following example:

```plaintext
proc univariate data=measures;
   ppplot width / beta(theta=1 sigma=2 alpha=3 beta=4);
run;
```

If you do not specify values for these parameters, then by default, \( \theta = 0, \sigma = 1, \) and maximum likelihood estimates are calculated for \( \alpha \) and \( \beta \).

**IMPORTANT:** If the default unit interval \((0,1)\) does not adequately describe the range of your data, then you should specify THETA=\(\theta\) and SIGMA=\(\sigma\) so that your data fall in the interval \((\theta, \theta + \sigma)\).

If the data are beta distributed with parameters \( \alpha, \beta, \sigma, \) and \( \theta \), then the points on the plot for ALPHA=\(\alpha\), BETA=\(\beta\), SIGMA=\(\sigma\), and THETA=\(\theta\) tend to fall on or near the diagonal line \( y = x \), which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified beta distribution is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.
BETA=value
specifies the shape parameter $\beta$ ($\beta > 0$) for P-P plots requested with the BETA distribution option. See the preceding entry for the BETA distribution option for an example.

C=value
specifies the shape parameter $c$ ($c > 0$) for P-P plots requested with the WEIBULL option. See the entry for the WEIBULL option for examples.

EXPO_NENTIAL< (exponential-options) >
EXP< (exponential-options) >
creates an exponential P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The $y$-coordinate of the $i$th point is the empirical cdf value $\hat{F}(X_i)$. The $x$-coordinate is the theoretical exponential cdf value

$$F(x_{(i)}) = 1 - \exp\left(-\frac{x_{(i)} - \theta}{\sigma}\right)$$

where

- $\theta$ = threshold parameter
- $\sigma$ = scale parameter ($\sigma > 0$)

You can specify $\sigma$ and $\theta$ with the SIGMA= and THETA= exponential-options, as illustrated in the following example:

```plaintext
proc univariate data=measures;
    ppplot width / exponential(theta=1 sigma=2);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and a maximum likelihood estimate is calculated for $\sigma$.

IMPORTANT: Your data must be greater than or equal to the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, specify $\theta$ with the THETA= option.

If the data are exponentially distributed with parameters $\sigma$ and $\theta$, the points on the plot for SIGMA=$\sigma$ and THETA=$\theta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified exponential distribution is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

GAMMA< (gamma-options) >
creates a gamma P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$
The $y$-coordinate of the $i$th point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical gamma cdf value

$$G_\alpha \left( \frac{x(i) - \theta}{\sigma} \right) = \int_{\theta}^{x(i)} \frac{1}{\sigma \Gamma(\alpha)} \left( \frac{t - \theta}{\sigma} \right)^{\alpha-1} \exp \left( -\frac{t - \theta}{\sigma} \right) dt$$

where $G_\alpha(\cdot)$ is the normalized incomplete gamma function and

$$\theta = \text{threshold parameter}$$
$$\sigma = \text{scale parameter (} \sigma > 0)$$
$$\alpha = \text{shape parameter (} \alpha > 0)$$

You can specify $\alpha$, $\sigma$, and $\theta$ with the ALPHA=, SIGMA=, and THETA= gamma-options, as illustrated in the following example:

```plaintext
proc univariate data=measures;
  ppplot width / gamma(alpha=1 sigma=2 theta=3);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and maximum likelihood estimates are calculated for $\alpha$ and $\sigma$.

**IMPORTANT:** Your data must be greater than or equal to the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, specify $\theta$ with the THETA= option.

If the data are gamma distributed with parameters $\alpha$, $\sigma$, and $\theta$, the points on the plot for ALPHA=$\alpha$, SIGMA=$\sigma$, and THETA=$\theta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified gamma distribution is a good fit. You can specify the SHAPE= option as an alias for the ALPHA= option, the SCALE= option as an alias for the SIGMA= option, and the THRESHOLD= option as an alias for the THETA= option.

---

**GUMBEL< (Gumbel-options) >**

creates a Gumbel P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

$$x(1) \leq x(2) \leq \cdots \leq x(n)$$

The $y$-coordinate of the $i$th point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical Gumbel cdf value

$$F(x(i)) = \exp(-e^{-(x(i) - \mu)/\sigma})$$

where

$$\mu = \text{location parameter}$$
$$\sigma = \text{scale parameter (} \sigma > 0)$$

You can specify $\mu$ and $\sigma$ with the MU= and SIGMA= Gumbel-options, as illustrated in the following example:
**Chapter 4: The UNIVARIATE Procedure**

```plaintext
proc univariate data=measures;
   ppplot width / gumbel(mu=1 sigma=2);
run;
```

If you do not specify values for these parameters, then by default, the maximum likelihood estimates are calculated for \( \mu \) and \( \sigma \).

If the data are Gumbel distributed with parameters \( \mu \) and \( \sigma \), the points on the plot for \( \text{MU}=\mu \) and \( \text{SIGMA}=\sigma \) tend to fall on or near the diagonal line \( y = x \), which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Gumbel distribution is a good fit.

**IGAUSS\((i\text{Gauss-options})\)**

creates an inverse Gaussian P-P plot. To create the plot, the \( n \) nonmissing observations are ordered from smallest to largest:

\[
x_1 \leq x_2 \leq \cdots \leq x_n
\]

The \( y \)-coordinate of the \( i \)th point is the empirical cdf value \( \frac{i}{n} \). The \( x \)-coordinate is the theoretical inverse Gaussian cdf value

\[
F(x_i) = \Phi \left\{ \sqrt{\frac{\lambda}{x_i}} \left( \frac{x_i}{\mu} - 1 \right) \right\} + e^{2\lambda/\mu} \Phi \left\{ -\sqrt{\frac{\lambda}{x_i}} \left( \frac{x_i}{\mu} + 1 \right) \right\}
\]

where \( \Phi() \) is the standard normal distribution function and

- \( \mu = \text{mean parameter} \ (\mu > 0) \)
- \( \lambda = \text{shape parameter} \ (\lambda > 0) \)

You can specify \( \lambda \) and \( \mu \) with the LAMBDA= and MU= IGAUSS-options, as illustrated in the following example:

```plaintext
proc univariate data=measures;
   ppplot width / igauss(lambda=1 mu=2);
run;
```

If you do not specify values for these parameters, then by default, the maximum likelihood estimates are calculated for \( \lambda \) and \( \mu \).

If the data are inverse Gaussian distributed with parameters \( \lambda \) and \( \mu \), the points on the plot for \( \text{LAMBDA}=\lambda \) and \( \text{MU}=\mu \) tend to fall on or near the diagonal line \( y = x \), which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified inverse Gaussian distribution is a good fit.

**LAMBDA=value**

specifies the shape parameter \( \lambda \) for fitted curves requested with the IGAUSS option. Enclose the LAMBDA= option in parentheses after the IGAUSS distribution keyword. If you do not specify a value for \( \lambda \), the procedure calculates a maximum likelihood estimate.
creates a lognormal P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

$$x(1) \leq x(2) \leq \cdots \leq x(n)$$

The $y$-coordinate of the $i$th point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical lognormal cdf value

$$\Phi\left(\frac{\log(x(i) - \theta) - \zeta}{\sigma}\right)$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function and

- $\theta =$ threshold parameter
- $\zeta =$ scale parameter
- $\sigma =$ shape parameter ($\sigma > 0$)

You can specify $\theta$, $\zeta$, and $\sigma$ with the THETA=, ZETA=, and SIGMA= lognormal-options, as illustrated in the following example:

```plaintext
proc univariate data=measures;
   ppplot width / lognormal(theta=1 zeta=2);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and maximum likelihood estimates are calculated for $\sigma$ and $\zeta$.

**IMPORTANT:** Your data must be greater than the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, specify $\theta$ with the THETA= option.

If the data are lognormally distributed with parameters $\sigma$, $\theta$, and $\zeta$, the points on the plot for SIGMA=$\sigma$, THETA=$\theta$, and ZETA=$\zeta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified lognormal distribution is a good fit. You can specify the SHAPE= option as an alias for the SIGMA=option, the SCALE= option as an alias for the ZETA= option, and the THRESHOLD= option as an alias for the THETA= option.

**MU=value**

specifies the parameter $\mu$ for P-P plots requested with the GUMBEL, IGAUSS, and NORMAL options. By default, the sample mean is used for $\mu$ with inverse Gaussian and normal distributions. A maximum likelihood estimate is computed by default with the Gumbel distribution. See Example 4.36.

**NOLINE**

suppresses the diagonal reference line.
**NORMAL< (normal-options ) >**

**NORM< (normal-options ) >**

creates a normal P-P plot. By default, if you do not specify a distribution option, the procedure displays a normal P-P plot. To create the plot, the \( n \) nonmissing observations are ordered from smallest to largest:

\[
x_1 \leq x_2 \leq \cdots \leq x_n
\]

The \( y \)-coordinate of the \( i \)th point is the empirical cdf value \( \frac{i}{n} \). The \( x \)-coordinate is the theoretical normal cdf value

\[
\Phi\left(\frac{x_i - \mu}{\sigma}\right) = \int_{-\infty}^{x_i} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(t - \mu)^2}{2\sigma^2}\right) dt
\]

where \( \Phi(\cdot) \) is the cumulative standard normal distribution function and

- \( \mu = \) location parameter or mean
- \( \sigma = \) scale parameter or standard deviation (\( \sigma > 0 \))

You can specify \( \mu \) and \( \sigma \) with the MU= and SIGMA= **normal-options**, as illustrated in the following example:

```plaintext
proc univariate data=measures;
  ppplot width / normal(mu=1 sigma=2);
run;
```

By default, the sample mean and sample standard deviation are used for \( \mu \) and \( \sigma \).

If the data are normally distributed with parameters \( \mu \) and \( \sigma \), the points on the plot for MU=\( \mu \) and SIGMA=\( \sigma \) tend to fall on or near the diagonal line \( y = x \), which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified normal distribution is a good fit. See Example 4.36.

**PARETO< (Pareto-options) >**

creates a generalized Pareto P-P plot. To create the plot, the \( n \) nonmissing observations are ordered from smallest to largest:

\[
x_1 \leq x_2 \leq \cdots \leq x_n
\]

The \( y \)-coordinate of the \( i \)th point is the empirical cdf value \( \frac{i}{n} \). The \( x \)-coordinate is the theoretical generalized Pareto cdf value

\[
F(x_{(i)}) = 1 - \left(1 - \frac{\alpha(x_{(i)} - \theta)}{\sigma}\right)^\frac{1}{\sigma}
\]

where

- \( \theta = \) threshold parameter
- \( \sigma = \) scale parameter (\( \sigma > 0 \))
- \( \alpha = \) shape parameter
The parameter $\theta$ for the generalized Pareto distribution must be less than the minimum data value. You can specify $\theta$ with the THETA= Pareto-option. The default value for $\theta$ is 0. In addition, the generalized Pareto distribution has a shape parameter $\alpha$ and a scale parameter $\sigma$. You can specify these parameters with the ALPHA= and SIGMA= Pareto-options. By default, maximum likelihood estimates are computed for $\alpha$ and $\sigma$.

If the data are generalized Pareto distributed with parameters $\theta$, $\sigma$, and $\alpha$, the points on the plot for THETA=$\theta$, SIGMA=$\sigma$, and ALPHA=$\alpha$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified generalized Pareto distribution is a good fit.

**POWER< (Power-options) >**

creates a power function P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The $y$-coordinate of the $i$th point is the empirical cdf value $i/n$. The $x$-coordinate is the theoretical power function cdf value

$$F(x_{(i)}) = \left( \frac{x_{(i)} - \theta}{\sigma} \right)^\alpha$$

where

$\theta = \text{lower threshold parameter (lower endpoint)}$

$\sigma = \text{scale parameter ($\sigma > 0$)}$

$\alpha = \text{shape parameter ($\alpha > 0$)}$

The power function distribution is bounded below by the parameter $\theta$ and above by the value $\theta + \sigma$. You can specify $\theta$ and $\sigma$ by using the THETA= and SIGMA= power-options. The default values for $\theta$ and $\sigma$ are 0 and 1, respectively.

You can specify a value for the shape parameter, $\alpha$, with the ALPHA= power-option. If you do not specify a value for $\alpha$, the procedure calculates a maximum likelihood estimate.

The power function distribution is a special case of the beta distribution with its second shape parameter, $\beta = 1$.

If the data are power function distributed with parameters $\theta$, $\sigma$, and $\alpha$, the points on the plot for THETA=$\theta$, SIGMA=$\sigma$, and ALPHA=$\alpha$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified power function distribution is a good fit.

**RAYLEIGH< (Rayleigh-options) >**

creates a Rayleigh P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The $y$-coordinate of the $i$th point is the empirical cdf value $i/n$. The $x$-coordinate is the theoretical Rayleigh cdf value

$$F(x_{(i)}) = 1 - e^{-(x_{(i)}-\theta)^2/(2\sigma^2)}$$
where

\[ \theta = \text{threshold parameter} \]
\[ \sigma = \text{scale parameter } (\sigma > 0) \]

The parameter \( \theta \) for the Rayleigh distribution must be less than the minimum data value. You can specify \( \theta \) with the \texttt{THETA= Rayleigh-option}. The default value for \( \theta \) is 0. You can specify \( \sigma \) with the \texttt{SIGMA= Rayleigh-option}. By default, a maximum likelihood estimate is computed for \( \sigma \).

If the data are Rayleigh distributed with parameters \( \theta \) and \( \sigma \), the points on the plot for \texttt{THETA=}\( \theta \) and \texttt{SIGMA=}\( \sigma \) tend to fall on or near the diagonal line \( y = x \), which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Rayleigh distribution is a good fit.

\texttt{SIGMA=value}

specifies the parameter \( \sigma \), where \( \sigma > 0 \). When used with the \texttt{BETA}, \texttt{EXPONENTIAL}, \texttt{GAMMA}, \texttt{GUMBEL}, \texttt{NORMAL}, \texttt{PARETO}, \texttt{POWER}, \texttt{RAYLEIGH}, and \texttt{WEIBULL} options, the \texttt{SIGMA=} option specifies the scale parameter. When used with the \texttt{LOGNORMAL} option, the \texttt{SIGMA=} option specifies the shape parameter. See Example 4.36.

\texttt{SQUARE}

displays the P-P plot in a square frame. The default is a rectangular frame. See Example 4.36.

\texttt{THETA=value}

\texttt{THRESHOLD=value}

specifies the lower threshold parameter \( \theta \) for plots requested with the \texttt{BETA}, \texttt{EXPONENTIAL}, \texttt{GAMMA}, \texttt{LOGNORMAL}, \texttt{PARETO}, \texttt{POWER}, \texttt{RAYLEIGH}, and \texttt{WEIBULL} options.

\texttt{WEIBULL< (Weibull-options) >}

\texttt{WEIB< (Weibull-options) >}

creates a Weibull P-P plot. To create the plot, the \( n \) nonmissing observations are ordered from smallest to largest:

\[ x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \]

The \( y \)-coordinate of the \( i \)th point is the empirical cdf value \( \frac{i}{n} \). The \( x \)-coordinate is the theoretical Weibull cdf value

\[ F(x_{(i)}) = 1 - \exp \left( - \left( \frac{x_{(i)} - \theta}{\sigma} \right)^{c} \right) \]

where

\[ \theta = \text{threshold parameter} \]
\[ \sigma = \text{scale parameter } (\sigma > 0) \]
\[ c = \text{shape parameter } (c > 0) \]

You can specify \( c \), \( \sigma \), and \( \theta \) with the \texttt{C=}, \texttt{SIGMA=}, and \texttt{THETA=} \texttt{Weibull-options}, as illustrated in the following example:
The `PROC UNIVARIATE` procedure is used to analyze the distribution of a variable. The `PROBPLOT` statement is used to create probability plots, which compare the observed distribution of a variable with a theoretical distribution. The `WEIBULL` distribution is specified in the `PROBPLOT` statement, with default parameters `C=1` and `SIGMA=2`. If you do not specify values for these parameters, then by default `C=0` and maximum likelihood estimates are calculated for `SIGMA` and `C`.

**IMPORTANT:** Your data must be greater than or equal to the lower threshold. If the default `C=0` is not an adequate lower bound for your data, you should specify `C` with the `THETA=` option.

If the data are Weibull distributed with parameters `C`, `SIGMA`, and `THETA`, the points on the plot for `C=C`, `SIGMA=SIGMA`, and `THETA=THETA` tend to fall on or near the diagonal line `y=x`, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Weibull distribution is a good fit. You can specify the `SHAPE=` option as an alias for the `C=` option, the `SCALE=` option as an alias for the `SIGMA=` option, and the `THRESHOLD=` option as an alias for the `THETA=` option.

**ZETA=value**

specifies a value for the scale parameter `zeta` for lognormal P-P plots requested with the LOGNORMAL option.

**PROBPLOT Statement**

```
PROBPLOT < variables > < / options > ;
```

The `PROBPLOT` statement creates a probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution. If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern. Consequently, you can use a probability plot to determine how well a theoretical distribution models a set of measurements.

Probability plots are similar to Q-Q plots, which you can create with the `QQPLOT` statement. Probability plots are preferable for graphical estimation of percentiles, whereas Q-Q plots are preferable for graphical estimation of distribution parameters.

You can use any number of `PROBPLOT` statements in the `UNIVARIATE` procedure. The components of the `PROBPLOT` statement are as follows.

**variables**

are the variables for which probability plots are created. If you specify a `VAR` statement, the `variables` must also be listed in the `VAR` statement. Otherwise, the `variables` can be any numeric variables in the input data set. If you do not specify a list of `variables`, then by default the procedure creates a probability plot for each variable listed in the `VAR` statement, or for each numeric variable in the `DATA=` data set if you do not specify a `VAR` statement. For example, each of the following `PROBPLOT` statements produces two probability plots, one for `Length` and one for `Width`:

```
proc univariate data=Measures;
    var Length Width;
    probplot;
run;
```

```
proc univariate data=Measures;
run;
```
PROC UNIVARIATE DATA=Measures;
   PROBPLOT Length1 Length2 / normal(mu=10 sigma=0.3)
                      square ctext=blue;
RUN;

**Table 4.18 through Table 4.20** list the PROBPLOT **options** by function. For complete descriptions, see the sections “Dictionary of Options” on page 350 and “Dictionary of Common Options” on page 369. **Options** can be any of the following:

- primary options
- secondary options
- general options

**Distribution Options**

**Table 4.18** lists **options** for requesting a theoretical distribution.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA(beta-options)</td>
<td>specifies beta probability plot for shape parameters ( \alpha ) and ( \beta ) specified with mandatory ALPHA= and BETA= beta-options</td>
</tr>
<tr>
<td>EXPONENTIAL(exponential-options)</td>
<td>specifies exponential probability plot</td>
</tr>
<tr>
<td>GAMMA(gamma-options)</td>
<td>specifies gamma probability plot for shape parameter ( \alpha ) specified with mandatory ALPHA= gamma-option</td>
</tr>
<tr>
<td>GUMBEL(Gumbel-options)</td>
<td>specifies Gumbel probability plot</td>
</tr>
</tbody>
</table>
Table 4.18 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOGNORMAL(lognormal-options)</td>
<td>specifies lognormal probability plot for shape parameter ( \sigma ) specified with mandatory SIGMA= lognormal-option</td>
</tr>
<tr>
<td>NORMAL(normal-options)</td>
<td>specifies normal probability plot</td>
</tr>
<tr>
<td>PARETO(Pareto-options)</td>
<td>specifies generalized Pareto probability plot for shape parameter ( \alpha ) specified with mandatory ALPHA= Pareto-option</td>
</tr>
<tr>
<td>POWER(power-options)</td>
<td>specifies power function probability plot for shape parameter ( \alpha ) specified with mandatory ALPHA= power-option</td>
</tr>
<tr>
<td>RAYLEIGH(Rayleigh-options)</td>
<td>specifies Rayleigh probability plot</td>
</tr>
<tr>
<td>WEIBULL(Weibull-options)</td>
<td>specifies three-parameter Weibull probability plot for shape parameter ( c ) specified with mandatory C= Weibull-option</td>
</tr>
<tr>
<td>WEIBULL2(Weibull2-options)</td>
<td>specifies two-parameter Weibull probability plot</td>
</tr>
</tbody>
</table>

Table 4.19 lists secondary options that specify distribution parameters and control the display of a distribution reference line. Specify these options in parentheses after the distribution keyword. For example, you can request a normal probability plot with a distribution reference line by specifying the NORMAL option as follows:

```plaintext
proc univariate;
    probplot Length / normal(mu=10 sigma=0.3 color=red);
run;
```

The MU= and SIGMA= normal-options display a distribution reference line that corresponds to the normal distribution with mean \( \mu_0 = 10 \) and standard deviation \( \sigma_0 = 0.3 \), and the COLOR= normal-option specifies the color for the line.

Table 4.19 Secondary Distribution Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR=</td>
<td>specifies color of distribution reference line</td>
</tr>
<tr>
<td>L=</td>
<td>specifies line type of distribution reference line</td>
</tr>
<tr>
<td>W=</td>
<td>specifies width of distribution reference line</td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter ( \alpha )</td>
</tr>
<tr>
<td>BETA=</td>
<td>specifies mandatory shape parameter ( \beta )</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies ( \sigma_0 ) for distribution reference line</td>
</tr>
</tbody>
</table>
Table 4.19  (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>THETA=</strong></td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Exponential-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Gamma-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter $\alpha$</td>
</tr>
<tr>
<td>ALPHADELTA=</td>
<td>specifies change in successive estimates of $\alpha$ at which the Newton-</td>
</tr>
<tr>
<td></td>
<td>Raphson approximation of $\hat{\alpha}$ terminates</td>
</tr>
<tr>
<td>ALPHAINITIAL=</td>
<td>specifies initial value for $\alpha$ in the Newton-Raphson approximation of</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}$</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approx-</td>
</tr>
<tr>
<td></td>
<td>imation of $\hat{\alpha}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Gumbel-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies $\mu_0$ for distribution reference line</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Lognormal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies mandatory shape parameter $\sigma$</td>
</tr>
<tr>
<td>SLOPE=</td>
<td>specifies slope of distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td>ZETA=</td>
<td>specifies $\zeta_0$ for distribution reference line (slope is $\exp(\zeta_0)$)</td>
</tr>
<tr>
<td><strong>Normal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies $\mu_0$ for distribution reference line</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Pareto-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter $\alpha$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Power-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter $\alpha$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Rayleigh-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Weibull-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies mandatory shape parameter $c$</td>
</tr>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approx-</td>
</tr>
<tr>
<td></td>
<td>imation of $\hat{c}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Weibull2-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies $c_0$ for distribution reference line (slope is $1/c_0$)</td>
</tr>
</tbody>
</table>
### Table 4.19  (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approx-</td>
</tr>
<tr>
<td></td>
<td>imation of $\hat{c}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line (intercept is $\log(\sigma_0)$)</td>
</tr>
<tr>
<td>SLOPE=</td>
<td>specifies slope of distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies known lower threshold $\theta_0$</td>
</tr>
</tbody>
</table>

### General Graphics Options

Table 4.20 summarizes the general options for enhancing probability plots.

#### Table 4.20  General Graphics Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Graphics Options</td>
<td></td>
</tr>
<tr>
<td>GRID</td>
<td>creates a grid</td>
</tr>
<tr>
<td>HREF=</td>
<td>specifies reference lines perpendicular to the horizontal axis</td>
</tr>
<tr>
<td>HREFLABELS=</td>
<td>specifies labels for HREF= lines</td>
</tr>
<tr>
<td>HREFLABPOS=</td>
<td>specifies position for HREF= line labels</td>
</tr>
<tr>
<td>NOHLABEL</td>
<td>suppresses label for horizontal axis</td>
</tr>
<tr>
<td>NOV LABEL</td>
<td>suppresses label for vertical axis</td>
</tr>
<tr>
<td>NOVTICK</td>
<td>suppresses tick marks and tick mark labels for vertical axis</td>
</tr>
<tr>
<td>PCTLORDER=</td>
<td>specifies tick mark labels for percentile axis</td>
</tr>
<tr>
<td>ROTATE</td>
<td>switches horizontal and vertical axes</td>
</tr>
<tr>
<td>SQUARE</td>
<td>displays plot in square format</td>
</tr>
<tr>
<td>VREF=</td>
<td>specifies reference lines perpendicular to the vertical axis</td>
</tr>
<tr>
<td>VREFLABELS=</td>
<td>specifies labels for VREF= lines</td>
</tr>
<tr>
<td>VREFLABPOS=</td>
<td>specifies horizontal position of labels for VREF= lines</td>
</tr>
<tr>
<td>VAXISLABEL=</td>
<td>specifies label for vertical axis</td>
</tr>
<tr>
<td>Options for Traditional Graphics Output</td>
<td></td>
</tr>
<tr>
<td>ANNOTATE=</td>
<td>specifies annotate data set</td>
</tr>
<tr>
<td>CAXIS=</td>
<td>specifies color for axis</td>
</tr>
<tr>
<td>CFRA ME=</td>
<td>specifies color for frame</td>
</tr>
<tr>
<td>CGRID=</td>
<td>specifies color for grid lines</td>
</tr>
<tr>
<td>CHREF=</td>
<td>specifies colors for HREF= lines</td>
</tr>
<tr>
<td>CSTATREF=</td>
<td>specifies colors for STATREF= lines</td>
</tr>
<tr>
<td>CTEXT=</td>
<td>specifies color for text</td>
</tr>
<tr>
<td>CVREF=</td>
<td>specifies colors for VREF= lines</td>
</tr>
<tr>
<td>DESCRIPTION=</td>
<td>specifies description for plot in graphics catalog</td>
</tr>
<tr>
<td>FONT=</td>
<td>specifies software font for text</td>
</tr>
<tr>
<td>HAXIS=</td>
<td>specifies AXIS statement for horizontal axis</td>
</tr>
<tr>
<td>HEIGHT=</td>
<td>specifies height of text used outside framed areas</td>
</tr>
<tr>
<td>HMINOR=</td>
<td>specifies number of horizontal minor tick marks</td>
</tr>
<tr>
<td>INFONT=</td>
<td>specifies software font for text inside framed areas</td>
</tr>
</tbody>
</table>
### Table 4.20 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>INHEIGHT=</td>
<td>specifies height of text inside framed areas</td>
</tr>
<tr>
<td>LGRID=</td>
<td>specifies a line type for grid lines</td>
</tr>
<tr>
<td>LHREF=</td>
<td>specifies line types for HREF= lines</td>
</tr>
<tr>
<td>LSTATREF=</td>
<td>specifies line types for STATREF= lines</td>
</tr>
<tr>
<td>LVREF=</td>
<td>specifies line types for VREF= lines</td>
</tr>
<tr>
<td>NAME=</td>
<td>specifies name for plot in graphics catalog</td>
</tr>
<tr>
<td>NOFRAME</td>
<td>suppresses frame around plotting area</td>
</tr>
<tr>
<td>PCTLMINOR</td>
<td>requests minor tick marks for percentile axis</td>
</tr>
<tr>
<td>WAXIS=</td>
<td>specifies line thickness for axes and frame</td>
</tr>
<tr>
<td>WGRID=</td>
<td>specifies line thickness for grid</td>
</tr>
<tr>
<td>TURNVLABELS</td>
<td>turns and vertically strings out characters in labels for vertical axis</td>
</tr>
<tr>
<td>VAXIS=</td>
<td>specifies AXIS statement for vertical axis</td>
</tr>
<tr>
<td>VMINOR=</td>
<td>specifies number of vertical minor tick marks</td>
</tr>
</tbody>
</table>

#### Options for ODS Graphics Output

| ODSFOOTNOTE= | specifies footnote displayed on plot                                        |
| ODSFOOTNOTE2=| specifies secondary footnote displayed on plot                              |
| ODDSTITLE=   | specifies title displayed on plot                                           |
| ODDSTITLE2=  | specifies secondary title displayed on plot                                 |
| OVERLAY      | overlays plots for different class levels (ODS Graphics only)               |

#### Options for Comparative Plots

| ANNOKEY      | applies annotation requested in ANNOTATE= data set to key cell only         |
| CFRAMESIDE=  | specifies color for filling frame for row labels                            |
| CFRAMETOP=   | specifies color for filling frame for column labels                         |
| CPROP=       | specifies color for proportion of frequency bar                            |
| CTEXTSIDE=   | specifies color for row labels                                             |
| CTEXTTOP=    | specifies color for column labels                                          |
| INTERTILE=   | specifies distance between tiles                                           |
| NCOLS=       | specifies number of columns in comparative probability plot                |
| NROWS=       | specifies number of rows in comparative probability plot                   |

#### Miscellaneous Options

| CONTENTS=    | specifies table of contents entry for probability plot grouping            |
| NADJ=        | adjusts sample size when computing percentiles                             |
| RANKADJ=     | adjusts ranks when computing percentiles                                   |

### Dictionary of Options

The following entries provide detailed descriptions of options in the PROBPLOT statement. Options marked with † are applicable only when traditional graphics are produced. See the section “Dictionary of Common Options” on page 369 for detailed descriptions of options common to all plot statements.

**ALPHA=value-list | EST**

specifies the mandatory shape parameter $\alpha$ for probability plots requested with the BETA, GAMMA, PARETO, and POWER options. Enclose the ALPHA= option in parentheses after the distribution keyword. If you specify ALPHA=EST, a maximum likelihood estimate is computed for $\alpha$. 

creates a beta probability plot for each combination of the required shape parameters $\alpha$ and $\beta$ specified by the required ALPHA= and BETA= beta-options. If you specify ALPHA=EST and BETA=EST, the procedure creates a plot based on maximum likelihood estimates for $\alpha$ and $\beta$. You can specify the SCALE= beta-option as an alias for the SIGMA= beta-option and the THRESHOLD= beta-option as an alias for the THETA= beta-option. To create a plot that is based on maximum likelihood estimates for $\alpha$ and $\beta$, specify ALPHA=EST and BETA=EST.

To obtain graphical estimates of $\alpha$ and $\beta$, specify lists of values in the ALPHA= and BETA= beta-options, and select the combination of $\alpha$ and $\beta$ that most nearly linearizes the point pattern. To assess the point pattern, you can add a diagonal distribution reference line corresponding to lower threshold parameter $\theta_0$ and scale parameter $\sigma_0$ with the THETA= and SIGMA= beta-options. Alternatively, you can add a line that corresponds to estimated values of $\theta_0$ and $\sigma_0$ with the beta-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the beta distribution with parameters $\alpha$, $\beta$, $\theta_0$, and $\sigma_0$ is a good fit.

BETA=value-list | EST

B=value-list | EST

specifies the mandatory shape parameter $\beta$ for probability plots requested with the BETA option. Enclose the BETA= option in parentheses after the BETA option. If you specify BETA=EST, a maximum likelihood estimate is computed for $\beta$.

C=value-list | EST

specifies the shape parameter $c$ for probability plots requested with the WEIBULL and WEIBULL2 options. Enclose this option in parentheses after the WEIBULL or WEIBULL2 option. C= is a required Weibull-option in the WEIBULL option; in this situation, it accepts a list of values, or if you specify C=EST, a maximum likelihood estimate is computed for $c$. You can optionally specify C=value or C=EST as a Weibull2-option with the WEIBULL2 option to request a distribution reference line; in this situation, you must also specify Weibull2-option SIGMA=value or SIGMA=EST.

† CGRID=color

specifies the color for grid lines when a grid displays on the plot. This option also produces a grid.

EXPONENTIAL< (exponential-options) >

creates an exponential probability plot. To assess the point pattern, add a diagonal distribution reference line corresponding to $\theta_0$ and $\sigma_0$ with the THETA= and SIGMA= exponential-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter $\theta_0$ and the scale parameter $\sigma$ with the exponential-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters $\theta_0$ and $\sigma_0$ is a good fit. You can specify the SCALE= exponential-option as an alias for the SIGMA= exponential-option and the THRESHOLD= exponential-option as an alias for the THETA= exponential-option.

GAMMA(ALPHA=value | EST < gamma-options >)

creates a gamma probability plot for each value of the shape parameter $\alpha$ given by the mandatory ALPHA= gamma-option. If you specify ALPHA=EST, the procedure creates a plot based on a maximum likelihood estimate for $\alpha$. To obtain a graphical estimate of $\alpha$, specify a list of values for the ALPHA= gamma-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to $\theta_0$ and $\sigma_0$ with the THETA= and SIGMA= gamma-options. Alternatively, you can add a line corresponding to estimated values values
of the threshold parameter \( \theta_0 \) and the scale parameter \( \sigma \) with the \textit{gamma-options} THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the gamma distribution with parameters \( \alpha, \theta_0, \) and \( \sigma_0 \) is a good fit. You can specify the SCALE= \textit{gamma-option} as an alias for the SIGMA= \textit{gamma-option} and the THRESHOLD= \textit{gamma-option} as an alias for the THETA= \textit{gamma-option}.

**GRID**

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

**GUMBEL**

creates a Gumbel probability plot. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \mu_0 \) and \( \sigma_0 \) with the MU= and SIGMA= \textit{Gumbel-options}. Alternatively, you can add a line corresponding to estimated values of the location parameter \( \mu_0 \) and the scale parameter \( \sigma \) with the \textit{Gumbel-options} MU=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters \( \mu_0 \) and \( \sigma_0 \) is a good fit.

† **LGRID=** \textit{linetype}

specifies the line type for the grid requested by the GRID= option. By default, LGRID=1, which produces a solid line.

**LOGNORMAL**

creates a lognormal probability plot for each value of the shape parameter \( \sigma \) given by the mandatory SIGMA= \textit{lognormal-option}. If you specify SIGMA=EST, the procedure creates a plot based on a maximum likelihood estimate for \( \sigma \). To obtain a graphical estimate of \( \sigma \), specify a list of values for the SIGMA= \textit{lognormal-option} and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \zeta_0 \) with the THETA= and ZETA= \textit{lognormal-options}. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \( \theta_0 \) and the scale parameter \( \zeta_0 \) with the \textit{lognormal-options} THETA=EST and ZETA=EST. Agreement between the reference line and the point pattern indicates that the lognormal distribution with parameters \( \sigma, \theta_0, \) and \( \zeta_0 \) is a good fit. You can specify the THRESHOLD= \textit{lognormal-option} as an alias for the THETA= \textit{lognormal-option} and the SCALE= \textit{lognormal-option} as an alias for the ZETA= \textit{lognormal-option}. See Example 4.26.

**NORMAL**

creates a normal probability plot. This is the default if you omit a distribution option. To assess the point pattern, you can add a diagonal distribution reference line corresponding to \( \mu_0 \) and \( \sigma_0 \) with the MU= and SIGMA= \textit{normal-options}. Alternatively, you can add a line corresponding to estimated values of \( \mu_0 \) and \( \sigma_0 \) with the \textit{normal-options} MU=EST and SIGMA=EST; the estimates of the mean

**MU=** \textit{value} | EST

specifies the mean \( \mu_0 \) for a probability plot requested with the GUMBEL and NORMAL options. Enclose MU= in parentheses after the distribution keyword. You can specify MU=EST to request a distribution reference line with \( \mu_0 \) equal to the sample mean with the normal distribution. If you specify MU=EST for the Gumbel distribution, the procedure computes a maximum likelihood estimate.

**NADJ=** \textit{value}

specifies the adjustment value added to the sample size in the calculation of theoretical percentiles. By default, NADJ=\( \frac{1}{4} \). Refer to Chambers et al. (1983).

**NORMAL**

creates a normal probability plot. This is the default if you omit a distribution option. To assess the point pattern, you can add a diagonal distribution reference line corresponding to \( \mu_0 \) and \( \sigma_0 \) with the MU= and SIGMA= \textit{normal-options}. Alternatively, you can add a line corresponding to estimated values of \( \mu_0 \) and \( \sigma_0 \) with the \textit{normal-options} MU=EST and SIGMA=EST; the estimates of the mean
\( \mu_0 \) and the standard deviation \( \sigma_0 \) are the sample mean and sample standard deviation. Agreement between the reference line and the point pattern indicates that the normal distribution with parameters \( \mu_0 \) and \( \sigma_0 \) is a good fit.

**PARETO(ALPHA=value | EST < Pareto-options>)**

creates a generalized Pareto probability plot for each value of the shape parameter \( \alpha \) given by the mandatory ALPHA= Pareto-option. If you specify ALPHA=EST, the procedure creates a plot based on a maximum likelihood estimate for \( \alpha \). To obtain a graphical estimate of \( \alpha \), specify a list of values for the ALPHA= Pareto-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= Pareto-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \( \theta_0 \) and the scale parameter \( \sigma \) with the Pareto-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the generalized Pareto distribution with parameters \( \alpha \), \( \theta_0 \), and \( \sigma_0 \) is a good fit.

† **PCTLMINOR**

requests minor tick marks for the percentile axis. The HMINOR option overrides the minor tick marks requested by the PCTLMINOR option.

**PCTLORDER=values**

specifies the tick marks that are labeled on the theoretical percentile axis. Because the values are percentiles, the labels must be between 0 and 100, exclusive. The values must be listed in increasing order and must cover the plotted percentile range. Otherwise, the default values of 1, 5, 10, 25, 50, 75, 90, 95, and 99 are used.

**POWER(ALPHA=value | EST < power-options>)**

creates a power function probability plot for each value of the shape parameter \( \alpha \) given by the mandatory ALPHA= power-option. If you specify ALPHA=EST, the procedure creates a plot based on a maximum likelihood estimate for \( \alpha \). To obtain a graphical estimate of \( \alpha \), specify a list of values for the ALPHA= power-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= power-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \( \theta_0 \) and the scale parameter \( \sigma \) with the power-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the power function distribution with parameters \( \alpha \), \( \theta_0 \), and \( \sigma_0 \) is a good fit.

**RANKADJ=value**

specifies the adjustment value added to the ranks in the calculation of theoretical percentiles. By default, RANKADJ=\( -\frac{3}{8} \), as recommended by Blom (1958). Refer to Chambers et al. (1983) for additional information.

**RAYLEIGH< (Rayleigh-options) >**

creates an Rayleigh probability plot. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= Rayleigh-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \( \theta_0 \) and the scale parameter \( \sigma \) with the Rayleigh-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters \( \theta_0 \) and \( \sigma_0 \) is a good fit.
ROTATE

switches the horizontal and vertical axes so that the theoretical percentiles are plotted vertically while the data are plotted horizontally. Regardless of whether the plot has been rotated, horizontal axis options (such as HAXIS=) still refer to the horizontal axis, and vertical axis options (such as VAXIS=) still refer to the vertical axis. All other options that depend on axis placement adjust to the rotated axes.

SIGMA=value-list | EST

specifies the parameter $\sigma$, where $\sigma > 0$. Alternatively, you can specify SIGMA=EST to request a maximum likelihood estimate for $\sigma_0$. The interpretation and use of the SIGMA= option depend on the distribution option with which it is used. See Table 4.21 for a summary of how to use the SIGMA= option. You must enclose this option in parentheses after the distribution option.

Table 4.21  Uses of the SIGMA= Option

<table>
<thead>
<tr>
<th>Distribution Option</th>
<th>Use of the SIGMA= Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA</td>
<td>THETA=$\theta_0$ and SIGMA=$\sigma_0$ request a distribution reference line corresponding to $\theta_0$ and $\sigma_0$.</td>
</tr>
<tr>
<td>EXPONENTIAL</td>
<td></td>
</tr>
<tr>
<td>GAMMA</td>
<td></td>
</tr>
<tr>
<td>PARETO</td>
<td></td>
</tr>
<tr>
<td>POWER</td>
<td></td>
</tr>
<tr>
<td>RAYLEIGH</td>
<td></td>
</tr>
<tr>
<td>WEIBULL</td>
<td></td>
</tr>
<tr>
<td>GUMBEL</td>
<td>MU=$\mu_0$ and SIGMA=$\sigma_0$ request a distribution reference line corresponding to $\mu_0$ and $\sigma_0$.</td>
</tr>
<tr>
<td>LOGNORMAL</td>
<td>SIGMA=$\sigma_1 \ldots \sigma_n$ requests $n$ probability plots with shape parameters $\sigma_1 \ldots \sigma_n$. The SIGMA= option must be specified.</td>
</tr>
<tr>
<td>NORMAL</td>
<td>MU=$\mu_0$ and SIGMA=$\sigma_0$ request a distribution reference line corresponding to $\mu_0$ and $\sigma_0$. SIGMA=EST requests a line with $\sigma_0$ equal to the sample standard deviation.</td>
</tr>
<tr>
<td>WEIBULL2</td>
<td>SIGMA=$\sigma_0$ and C=$c_0$ request a distribution reference line corresponding to $\sigma_0$ and $c_0$.</td>
</tr>
</tbody>
</table>

SLOPE=value | EST

specifies the slope for a distribution reference line requested with the LOGNORMAL and WEIBULL2 options. Enclose the SLOPE= option in parentheses after the distribution option. When you use the SLOPE= lognormal-option with the LOGNORMAL option, you must also specify a threshold parameter value $\theta_0$ with the THETA= lognormal-option to request the line. The SLOPE= lognormal-option is an alternative to the ZETA= lognormal-option for specifying $\xi_0$, because the slope is equal to $\exp(\xi_0)$.

When you use the SLOPE= Weibull2-option with the WEIBULL2 option, you must also specify a scale parameter value $\sigma_0$ with the SIGMA= Weibull2-option to request the line. The SLOPE= Weibull2-option is an alternative to the C= Weibull2-option 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 and the third and fourth PROBPLOT statements produce the same probability plots:
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);
run;

SQUARE displays the probability plot in a square frame. By default, the plot is in a rectangular frame.

THETA=value | EST
THRESHOLD=value | EST
specifies the lower threshold parameter \( \theta \) for plots requested with the BETA, EXPONENTIAL, GAMMA, PARETO, POWER, RAYLEIGH, LOGNORMAL, WEIBULL, and WEIBULL2 options. Enclose the THETA= option in parentheses after a distribution option. When used with the WEIBULL2 option, the THETA= option specifies the known lower threshold \( \theta_0 \), for which the default is 0. When used with the other distribution options, the THETA= option specifies \( \theta_0 \) for a distribution reference line; alternatively in this situation, you can specify THETA=EST to request a maximum likelihood estimate for \( \theta_0 \). To request the line, you must also specify a scale parameter.

WEIBULL(C=value | EST < Weibull-options>)
WEIB(C=value | EST < Weibull-options>)
creates a three-parameter Weibull probability plot for each value of the required shape parameter \( c \) specified by the mandatory C= Weibull-option. To create a plot that is based on a maximum likelihood estimate for \( c \), specify C=EST. To obtain a graphical estimate of \( c \), specify a list of values in the C= Weibull-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= Weibull-options. Alternatively, you can add a line corresponding to estimated values of \( \theta_0 \) and \( \sigma_0 \) with the Weibull-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters \( c_0 \), \( \theta_0 \), and \( \sigma_0 \) is a good fit. You can specify the SCALE= Weibull-option as an alias for the SIGMA= Weibull-option and the THRESHOLD= Weibull-option as an alias for the THETA= Weibull-option.

WEIBULL2< (Weibull2-options) >
W2< (Weibull2-options) >
creates a two-parameter Weibull probability plot. You should use the WEIBULL2 option when your data have a known lower threshold \( \theta_0 \), which is 0 by default. To specify the threshold value \( \theta_0 \), use the THETA= Weibull2-option. By default, THETA=0. An advantage of the two-parameter Weibull plot over the three-parameter Weibull plot is that the parameters \( c \) and \( \sigma \) 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. To obtain a graphical estimate of \( \theta_0 \), specify a list of values for the THETA= Weibull2-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \sigma_0 \) and \( c_0 \) with the SIGMA= and C= Weibull2-options. Alternatively, you can add a distribution reference line corresponding to estimated values of \( \sigma_0 \) and \( c_0 \) with the Weibull2-options SIGMA=EST and C=EST. Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters \( c_0 \), \( \theta_0 \), and \( \sigma_0 \) is a good fit. You can specify the SCALE= Weibull2-option
as an alias for the SIGMA=\textit{Weibull2-option} and the SHAPE=\textit{Weibull2-option} as an alias for the C=\textit{Weibull2-option}.

\textit{† WGRID=\textit{n}}
specifies the line thickness for the grid when producing traditional graphics. The option does not apply to ODS Graphics output.

\textit{ZETA=\textit{value} | \textit{EST}}
specifies a value for the scale parameter $\zeta$ for the lognormal probability plots requested with the LOGNORMAL option. Enclose the ZETA= lognormal-option in parentheses after the LOGNORMAL option. To request a distribution reference line with intercept $\theta_0$ and slope $\exp(\zeta_0)$, specify the THETA=$\theta_0$ and ZETA=$\zeta_0$.

**QQPLOT Statement**

```
QQPLOT < variables > < / options > ;
```

The QQPLOT statement creates quantile-quantile plots (Q-Q plots) and compares ordered variable values with quantiles of a specified theoretical distribution. 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 to determine how well a theoretical distribution models a set of measurements.

Q-Q plots are similar to probability plots, which you can create with the PROBPLOT statement. Q-Q plots are preferable for graphical estimation of distribution parameters, whereas probability plots are preferable for graphical estimation of percentiles.

You can use any number of QQPLOT statements in the UNIVARIATE procedure. The components of the QQPLOT statement are as follows.

\textit{variables}

are the variables for which Q-Q plots are created. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify a list of variables, then by default the procedure creates a Q-Q plot for each variable listed in the VAR statement, or for each numeric variable in the DATA= data set if you do not specify a VAR statement. For example, each of the following QQPLOT statements produces two Q-Q plots, one for Length and one for Width:

```
proc univariate data=Measures;
  var Length Width;
  qqqplot;

proc univariate data=Measures;
  qqqplot Length Width;
  run;
```

\textit{options}

specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all options after the slash (/) in the QQPLOT statement. You can specify only one \textit{option} that names the distribution in each QQPLOT
statement, but you can specify any number of other *options*. The distributions available are the beta, exponential, gamma, lognormal, normal, two-parameter Weibull, and three-parameter Weibull. By default, the procedure produces a plot for the normal distribution.

In the following example, the NORMAL option requests a normal Q-Q plot for each variable. The MU= and SIGMA= *normal-options* request a distribution reference line with intercept 10 and slope 0.3 for each plot, corresponding to a normal distribution with mean \( \mu = 10 \) and standard deviation \( \sigma = 0.3 \). The SQUARE option displays the plot in a square frame, and the CTEXT= option specifies the text color.

```plaintext
proc univariate data=measures;
    qplot length1 length2 / normal(mu=10 sigma=0.3)
         square ctext=blue;
run;
```

Table 4.22 through Table 4.24 list the QQPLOT *options* by function. For complete descriptions, see the sections “Dictionary of Options” on page 361 and “Dictionary of Common Options” on page 369.

*Options* can be any of the following:

- primary options
- secondary options
- general options

**Distribution Options**

Table 4.22 lists primary options for requesting a theoretical distribution. See the section “Distributions for Probability and Q-Q Plots” on page 421 for detailed descriptions of these distributions.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA(<em>beta-options</em>)</td>
<td>specifies beta Q-Q plot for shape parameters ( \alpha ) and ( \beta ) specified with mandatory ALFA= and BETA= <em>beta-options</em></td>
</tr>
<tr>
<td>EXPONENTIAL(<em>exponential-options</em>)</td>
<td>specifies exponential Q-Q plot</td>
</tr>
<tr>
<td>GAMMA(<em>gamma-options</em>)</td>
<td>specifies gamma Q-Q plot for shape parameter ( \alpha ) specified with mandatory ALFA= <em>gamma-option</em></td>
</tr>
<tr>
<td>GUMBEL(<em>Gumbel-options</em>)</td>
<td>specifies gumbel Q-Q plot</td>
</tr>
<tr>
<td>LOGNORMAL(<em>lognormal-options</em>)</td>
<td>specifies lognormal Q-Q plot for shape parameter ( \sigma ) specified with mandatory SIGMA= <em>lognormal-option</em></td>
</tr>
<tr>
<td>NORMAL(<em>normal-options</em>)</td>
<td>specifies normal Q-Q plot</td>
</tr>
<tr>
<td>PARETO(<em>Pareto-options</em>)</td>
<td>specifies generalized Pareto Q-Q plot for shape parameter ( \alpha ) specified with mandatory ALFA= <em>Pareto-option</em></td>
</tr>
</tbody>
</table>
Table 4.22  (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>POWER(power-options)</td>
<td>specifies power function Q-Q plot for shape parameter ( \alpha ) specified with mandatory ALPHA= power-option</td>
</tr>
<tr>
<td>RAYLEIGH(Rayleigh-options)</td>
<td>specifies Rayleigh Q-Q plot</td>
</tr>
<tr>
<td>WEIBULL(Weibull-options)</td>
<td>specifies three-parameter Weibull Q-Q plot for shape parameter ( c ) specified with mandatory C= Weibull-option</td>
</tr>
<tr>
<td>WEIBULL2(Weibull2-options)</td>
<td>specifies two-parameter Weibull Q-Q plot</td>
</tr>
</tbody>
</table>

Table 4.23 lists secondary options that specify distribution parameters and control the display of a distribution reference line. Specify these options in parentheses after the distribution keyword. For example, you can request a normal Q-Q plot with a distribution reference line by specifying the NORMAL option as follows:

```r
proc univariate;
  qqqplot Length / normal(mu=10 sigma=0.3 color=red);
run;
```

The MU= and SIGMA= normal-options display a distribution reference line that corresponds to the normal distribution with mean \( \mu_0 = 10 \) and standard deviation \( \sigma_0 = 0.3 \), and the COLOR= normal-option specifies the color for the line.

Table 4.23  Secondary Distribution Reference Line Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR=</td>
<td>specifies color of distribution reference line</td>
</tr>
<tr>
<td>L=</td>
<td>specifies line type of distribution reference line</td>
</tr>
<tr>
<td>W=</td>
<td>specifies width of distribution reference line</td>
</tr>
<tr>
<td>Beta-Options</td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter ( \alpha )</td>
</tr>
<tr>
<td>BETA=</td>
<td>specifies mandatory shape parameter ( \beta )</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies ( \sigma_0 ) for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies ( \theta_0 ) for distribution reference line</td>
</tr>
<tr>
<td>Exponential-Options</td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies ( \sigma_0 ) for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies ( \theta_0 ) for distribution reference line</td>
</tr>
<tr>
<td>Gamma-Options</td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter ( \alpha )</td>
</tr>
<tr>
<td>ALPHADELTA=</td>
<td>specifies change in successive estimates of ( \alpha ) at which the Newton-Raphson approximation of ( \hat{\alpha} ) terminates</td>
</tr>
<tr>
<td>ALPHAINITIAL=</td>
<td>specifies initial value for ( \alpha ) in the Newton-Raphson approximation of ( \hat{\alpha} )</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of ( \hat{\alpha} )</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies ( \sigma_0 ) for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies ( \theta_0 ) for distribution reference line</td>
</tr>
</tbody>
</table>
Table 4.23 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gumbel-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies $\mu_0$ for distribution reference line</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Lognormal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies mandatory shape parameter $\sigma$</td>
</tr>
<tr>
<td>SLOPE=</td>
<td>specifies slope of distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td>ZETA=</td>
<td>specifies $\zeta_0$ for distribution reference line (slope is $\exp(\zeta_0)$)</td>
</tr>
<tr>
<td><strong>Normal-Options</strong></td>
<td></td>
</tr>
<tr>
<td>MU=</td>
<td>specifies $\mu_0$ for distribution reference line</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Pareto-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter $\alpha$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Power-Options</strong></td>
<td></td>
</tr>
<tr>
<td>ALPHA=</td>
<td>specifies mandatory shape parameter $\alpha$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Rayleigh-Options</strong></td>
<td></td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Weibull-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies mandatory shape parameter $c$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Weibull2-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies $c_0$ for distribution reference line (slope is $1/c_0$)</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line (intercept is $\log(\sigma_0)$)</td>
</tr>
<tr>
<td>SLOPE=</td>
<td>specifies slope of distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies slope of distribution reference line</td>
</tr>
<tr>
<td><strong>Weibull-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies $\alpha$ for distribution reference line</td>
</tr>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{c}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies $\theta_0$ for distribution reference line</td>
</tr>
<tr>
<td><strong>Weibull2-Options</strong></td>
<td></td>
</tr>
<tr>
<td>C=</td>
<td>specifies $c_0$ for distribution reference line (slope is $1/c_0$)</td>
</tr>
<tr>
<td>ITPRINT</td>
<td>requests table of iteration history and optimizer details</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{c}$</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>specifies $\sigma_0$ for distribution reference line (intercept is $\log(\sigma_0)$)</td>
</tr>
</tbody>
</table>
### Table 4.23 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLOPE=</td>
<td>specifies slope of distribution reference line</td>
</tr>
<tr>
<td>THETA=</td>
<td>specifies known lower threshold $\theta_0$</td>
</tr>
</tbody>
</table>

### General Options

Table 4.24 summarizes general options for enhancing Q-Q plots.

### Table 4.24 General Graphics Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Graphics Options</strong></td>
<td></td>
</tr>
<tr>
<td>GRID</td>
<td>creates a grid</td>
</tr>
<tr>
<td>HREF=</td>
<td>specifies reference lines perpendicular to the horizontal axis</td>
</tr>
<tr>
<td>HREFLABELS=</td>
<td>specifies labels for HREF= lines</td>
</tr>
<tr>
<td>HREFLABPOS=</td>
<td>specifies vertical position of labels for HREF= lines</td>
</tr>
<tr>
<td>NOHLABEL</td>
<td>suppresses label for horizontal axis</td>
</tr>
<tr>
<td>NOVLABEL</td>
<td>suppresses label for vertical axis</td>
</tr>
<tr>
<td>NOVTICK</td>
<td>suppresses tick marks and tick mark labels for vertical axis</td>
</tr>
<tr>
<td>PCTLAXIS</td>
<td>displays a nonlinear percentile axis</td>
</tr>
<tr>
<td>PCTLScale</td>
<td>replaces theoretical quantiles with percentiles</td>
</tr>
<tr>
<td>ROTATE</td>
<td>switches horizontal and vertical axes</td>
</tr>
<tr>
<td>SQUARE</td>
<td>displays plot in square format</td>
</tr>
<tr>
<td>VAXISLABEL=</td>
<td>specifies label for vertical axis</td>
</tr>
<tr>
<td>VREF=</td>
<td>specifies reference lines perpendicular to the vertical axis</td>
</tr>
<tr>
<td>VREFLABELS=</td>
<td>specifies labels for VREF= lines</td>
</tr>
<tr>
<td>VREFLABPOS=</td>
<td>specifies horizontal position of labels for VREF= lines</td>
</tr>
</tbody>
</table>

**Options for Traditional Graphics Output**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNOTATE=</td>
<td>specifies annotate data set</td>
</tr>
<tr>
<td>CAXIS=</td>
<td>specifies color for axis</td>
</tr>
<tr>
<td>CFRAME=</td>
<td>specifies color for frame</td>
</tr>
<tr>
<td>CGRID=</td>
<td>specifies color for grid lines</td>
</tr>
<tr>
<td>CHREF=</td>
<td>specifies colors for HREF= lines</td>
</tr>
<tr>
<td>CSTATREF=</td>
<td>specifies colors for STATREF= lines</td>
</tr>
<tr>
<td>CTEXT=</td>
<td>specifies color for text</td>
</tr>
<tr>
<td>CVREF=</td>
<td>specifies colors for VREF= lines</td>
</tr>
<tr>
<td>DESCRIPTION=</td>
<td>specifies description for plot in graphics catalog</td>
</tr>
<tr>
<td>FONT=</td>
<td>specifies software font for text</td>
</tr>
<tr>
<td>HEIGHT=</td>
<td>specifies height of text used outside framed areas</td>
</tr>
<tr>
<td>HMINOR=</td>
<td>specifies number of horizontal minor tick marks</td>
</tr>
<tr>
<td>INFONT=</td>
<td>specifies software font for text inside framed areas</td>
</tr>
<tr>
<td>INHEIGHT=</td>
<td>specifies height of text inside framed areas</td>
</tr>
<tr>
<td>LGGRID=</td>
<td>specifies a line type for grid lines</td>
</tr>
<tr>
<td>LHREF=</td>
<td>specifies line types for HREF= lines</td>
</tr>
<tr>
<td>LSTATREF=</td>
<td>specifies line types for STATREF= lines</td>
</tr>
</tbody>
</table>
Table 4.24 (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVREF=</td>
<td>specifies line types for VREF= lines</td>
</tr>
<tr>
<td>NAME=</td>
<td>specifies name for plot in graphics catalog</td>
</tr>
<tr>
<td>NOFRAME</td>
<td>suppresses frame around plotting area</td>
</tr>
<tr>
<td>PCTLMINOR</td>
<td>requests minor tick marks for percentile axis</td>
</tr>
<tr>
<td>VAXIS=</td>
<td>specifies AXIS statement for vertical axis</td>
</tr>
<tr>
<td>VMINOR=</td>
<td>specifies number of vertical minor tick marks</td>
</tr>
<tr>
<td>WAXIS=</td>
<td>specifies line thickness for axes and frame</td>
</tr>
<tr>
<td>WGRID=</td>
<td>specifies line thickness for grid</td>
</tr>
</tbody>
</table>

Options for ODS Graphics Output

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODSFOOTNOTE=</td>
<td>specifies footnote displayed on plot</td>
</tr>
<tr>
<td>ODSFOOTNOTE2=</td>
<td>specifies secondary footnote displayed on plot</td>
</tr>
<tr>
<td>ODSTITLE=</td>
<td>specifies title displayed on plot</td>
</tr>
<tr>
<td>ODSTITLE2=</td>
<td>specifies secondary title displayed on plot</td>
</tr>
</tbody>
</table>

Options for Comparative Plots

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNOKEY</td>
<td>applies annotation requested in ANNOTATE= data set to key cell only</td>
</tr>
<tr>
<td>CFRAMESIDE=</td>
<td>specifies color for filling frame for row labels</td>
</tr>
<tr>
<td>CFRAMETOP=</td>
<td>specifies color for filling frame for column labels</td>
</tr>
<tr>
<td>CPROP=</td>
<td>specifies color for proportion of frequency bar</td>
</tr>
<tr>
<td>INTERTILE=</td>
<td>specifies distance between tiles</td>
</tr>
<tr>
<td>NCOLS=</td>
<td>specifies number of columns in comparative Q-Q plot</td>
</tr>
<tr>
<td>NROWS=</td>
<td>specifies number of rows in comparative Q-Q plot</td>
</tr>
</tbody>
</table>

Miscellaneous Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTENTS=</td>
<td>specifies table of contents entry for Q-Q plot grouping</td>
</tr>
<tr>
<td>NADJ=</td>
<td>adjusts sample size when computing percentiles</td>
</tr>
<tr>
<td>RANKADJ=</td>
<td>adjusts ranks when computing percentiles</td>
</tr>
</tbody>
</table>

Dictionary of Options

The following entries provide detailed descriptions of options in the QQPLOT statement. Options marked with † are applicable only when traditional graphics are produced. See the section “Dictionary of Common Options” on page 369 for detailed descriptions of options common to all plot statements.

**ALPHA=value-list | EST**

specifies the mandatory shape parameter $\alpha$ for quantile plots requested with the BETA, GAMMA, PARETO, and POWER options. Enclose the ALPHA= option in parentheses after the distribution keyword. If you specify ALPHA=EST, a maximum likelihood estimate is computed for $\alpha$.

**BETA(ALPHA=value | EST BETA=value | EST < beta-options>)**

creates a beta quantile plot for each combination of the required shape parameters $\alpha$ and $\beta$ specified by the required ALPHA= and BETA= beta-options. If you specify ALPHA=EST and BETA=EST, the procedure creates a plot based on maximum likelihood estimates for $\alpha$ and $\beta$. You can specify the SCALE= beta-option as an alias for the SIGMA= beta-option and the THRESHOLD= beta-option as an alias for the THETA= beta-option. To create a plot that is based on maximum likelihood estimates for $\alpha$ and $\beta$, specify ALPHA=EST and BETA=EST. See the section “Beta Distribution” on page 422 for details.
To obtain graphical estimates of \(\alpha\) and \(\beta\), specify lists of values in the ALPHA= and BETA= beta-options and select the combination of \(\alpha\) and \(\beta\) that most nearly linearizes the point pattern. To assess the point pattern, you can add a diagonal distribution reference line corresponding to lower threshold parameter \(\theta_0\) and scale parameter \(\sigma_0\) with the THETA= and SIGMA= beta-options. Alternatively, you can add a line that corresponds to estimated values of \(\theta_0\) and \(\sigma_0\) with the beta-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the beta distribution with parameters \(\alpha\), \(\beta\), \(\theta_0\), and \(\sigma_0\) is a good fit.

\[
\text{BETA=} \text{value-list} | \text{EST}
\]

specifies the mandatory shape parameter \(\beta\) for quantile plots requested with the BETA option. Enclose the BETA= option in parentheses after the BETA option. If you specify BETA=EST, a maximum likelihood estimate is computed for \(\beta\).

\[
\text{C=} \text{value-list} | \text{EST}
\]

specifies the shape parameter \(c\) for quantile plots requested with the WEIBULL and WEIBULL2 options. Enclose this option in parentheses after the WEIBULL or WEIBULL2 option. \(C=\) is a required Weibull-option in the WEIBULL option; in this situation, it accepts a list of values, or if you specify C=EST, a maximum likelihood estimate is computed for \(c\). You can optionally specify C= value or C=EST as a Weibull2-option with the WEIBULL2 option to request a distribution reference line; in this situation, you must also specify Weibull2-option SIGMA= value or SIGMA=EST.

\[
\dagger \text{GRID=} \text{color}
\]

specifies the color for grid lines when a grid displays on the plot. This option also produces a grid.

\[
\text{EXP} \langle \text{exponential-options} \rangle
\]

creates an exponential quantile plot. To assess the point pattern, add a diagonal distribution reference line corresponding to \(\theta_0\) and \(\sigma_0\) with the THETA= and SIGMA= exponential-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \(\theta_0\) and the scale parameter \(\sigma\) with the exponential-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters \(\theta_0\) and \(\sigma_0\) is a good fit. You can specify the SCALE= exponential-option as an alias for the SIGMA= exponential-option and the THRESHOLD= exponential-option as an alias for the THETA= exponential-option. See the section “Exponential Distribution” on page 423 for details.

\[
\text{GAMMA} \langle \text{ALPHA=} \text{value} | \text{EST} \langle \text{gamma-options} \rangle \rangle
\]

creates a gamma quantile plot for each value of the shape parameter \(\alpha\) given by the mandatory ALPHA= gamma-option. If you specify ALPHA=EST, the procedure creates a plot based on a maximum likelihood estimate for \(\alpha\). To obtain a graphical estimate of \(\alpha\), specify a list of values for the ALPHA= gamma-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \(\theta_0\) and \(\sigma_0\) with the THETA= and SIGMA= gamma-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \(\theta_0\) and the scale parameter \(\sigma\) with the gamma-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the gamma distribution with parameters \(\alpha\), \(\theta_0\), and \(\sigma_0\) is a good fit. You can specify the SCALE= gamma-option as an alias for the SIGMA= gamma-option and the THRESHOLD= gamma-option as an alias for the THETA= gamma-option. See the section “Gamma Distribution” on page 423 for details.
GRID displays a grid of horizontal lines positioned at major tick marks on the vertical axis.

GUMBEL<(Gumbel-options)> creates a Gumbel quantile plot. To assess the point pattern, add a diagonal distribution reference line corresponding to $\mu_0$ and $\sigma_0$ with the MU= and SIGMA= Gumbel-options. Alternatively, you can add a line corresponding to estimated values of the location parameter $\mu_0$ and the scale parameter $\sigma$ with the Gumbel-options MU=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters $\mu_0$ and $\sigma_0$ is a good fit. See the section “Gumbel Distribution” on page 424 for details.

† LGRID=linetype specifies the line type for the grid requested by the GRID option. By default, LGRID=1, which produces a solid line. The LGRID= option also produces a grid.

LOGNORMAL(SIGMA=value | EST < lognormal-options>) LNORM(SIGMA=value | EST < lognormal-options>) creates a lognormal quantile plot for each value of the shape parameter $\sigma$ given by the mandatory SIGMA= lognormal-option. If you specify SIGMA=EST, the procedure creates a plot based on a maximum likelihood estimate for $\sigma$. To obtain a graphical estimate of $\sigma$, specify a list of values for the SIGMA= lognormal-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to $\theta_0$ and $\zeta_0$ with the THETA= and ZETA= lognormal-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter $\theta_0$ and the scale parameter $\zeta_0$ with the lognormal-options THETA=EST and ZETA=EST. Agreement between the reference line and the point pattern indicates that the lognormal distribution with parameters $\sigma$, $\theta_0$, and $\zeta_0$ is a good fit. You can specify the THRESHOLD= lognormal-option as an alias for the THETA= lognormal-option and the SCALE= lognormal-option as an alias for the ZETA= lognormal-option. See the section “Lognormal Distribution” on page 424 for details, and see Example 4.31 through Example 4.33 for examples that use the LOGNORMAL option.

MU=value | EST specifies the mean $\mu_0$ for a quantile plot requested with the GUMBEL and NORMAL options. Enclose MU= in parentheses after the distribution keyword. You can specify MU=EST to request a distribution reference line with $\mu_0$ equal to the sample mean with the normal distribution. If you specify MU=EST for the Gumbel distribution, the procedure computes a maximum likelihood estimate.

NADJ=value specifies the adjustment value added to the sample size in the calculation of theoretical percentiles. By default, NADJ=\(\frac{1}{4}\). Refer to Chambers et al. (1983) for additional information.

NORMAL<(normal-options)> creates a normal quantile plot. This is the default if you omit a distribution option. To assess the point pattern, you can add a diagonal distribution reference line corresponding to $\mu_0$ and $\sigma_0$ with the MU= and SIGMA= normal-options. Alternatively, you can add a line corresponding to estimated values of $\mu_0$ and $\sigma_0$ with the normal-options MU=EST and SIGMA=EST; the estimates of the mean $\mu_0$ and the standard deviation $\sigma_0$ are the sample mean and sample standard deviation. Agreement between the reference line and the point pattern indicates that the normal distribution with parameters $\mu_0$ and $\sigma_0$ is a good fit. See the section “Normal Distribution” on page 424 for details, and see Example 4.28 and Example 4.30 for examples that use the NORMAL option.
Chapter 4: The UNIVARIATE Procedure

PARETO(ALPHA=value | EST < Pareto-options>)
creates a generalized Pareto quantile plot for each value of the shape parameter \( \alpha \) given by the mandatory ALPHA= Pareto-option. If you specify ALPHA=EST, the procedure creates a plot based on a maximum likelihood estimate for \( \alpha \). To obtain a graphical estimate of \( \alpha \), specify a list of values for the ALPHA= Pareto-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= Pareto-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \( \theta_0 \) and the scale parameter \( \sigma \) with the Pareto-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the generalized Pareto distribution with parameters \( \alpha \), \( \theta_0 \), and \( \sigma_0 \) is a good fit. See the section “Generalized Pareto Distribution” on page 425 for details.

PCTLAXIS< (axis-options) >
adds a nonlinear percentile axis along the frame of the Q-Q plot opposite the theoretical quantile axis. The added axis is identical to the axis for probability plots produced with the PROBPLOT statement. When using the PCTLAXIS option, you must specify HREF= values in quantile units, and you cannot use the NOFRAME option. You can specify the following axis-options:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGRID=</td>
<td>specifies color for grid lines</td>
</tr>
<tr>
<td>GRID</td>
<td>draws grid lines at major percentiles</td>
</tr>
<tr>
<td>LABEL='string'</td>
<td>specifies label for percentile axis</td>
</tr>
<tr>
<td>LGRID=linetype</td>
<td>specifies line type for grid</td>
</tr>
<tr>
<td>WGRID=n</td>
<td>specifies line thickness for grid</td>
</tr>
</tbody>
</table>

† PCTLMINOR
requests minor tick marks for the percentile axis when you specify PCTLAXIS. The HMINOR option overrides the PCTLMINOR option.

PCTLSCALE
requests scale labels for the theoretical quantile axis in percentile units, resulting in a nonlinear axis scale. 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 PCTLAXIS option or use the PROBPLOT statement.

POWER(ALPHA=value | EST < power-options>)
creates a power function quantile plot for each value of the shape parameter \( \alpha \) given by the mandatory ALPHA= power-option. If you specify ALPHA=EST, the procedure creates a plot based on a maximum likelihood estimate for \( \alpha \). To obtain a graphical estimate of \( \alpha \), specify a list of values for the ALPHA= power-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= power-options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter \( \theta_0 \) and the scale parameter \( \sigma \) with the power-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the power function distribution with parameters \( \alpha \), \( \theta_0 \), and \( \sigma_0 \) is a good fit. See the section “Power Function Distribution” on page 425 for details.
**QQPLOT Statement**

**RAYLEIGH< (Rayleigh-options) >**

creates a Rayleigh quantile plot. To assess the point pattern, add a diagonal distribution reference line corresponding to $\theta_0$ and $\sigma_0$ with the THETA= and SIGMA= options. Alternatively, you can add a line corresponding to estimated values of the threshold parameter $\theta_0$ and the scale parameter $\sigma$ with the THETA=EST and SIGMA=EST options. Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters $\theta_0$ and $\sigma_0$ is a good fit. See the section “Rayleigh Distribution” on page 425 for details.

**RANKADJ=value**

specifies the adjustment value added to the ranks in the calculation of theoretical percentiles. By default, RANKADJ=$-\frac{3}{8}$, as recommended by Blom (1958). Refer to Chambers et al. (1983) for additional information.

**ROTATE**

switches the horizontal and vertical axes so that the theoretical quantiles are plotted vertically while the data are plotted horizontally. Regardless of whether the plot has been rotated, horizontal axis options (such as HAXIS=) still refer to the horizontal axis, and vertical axis options (such as VAXIS=) still refer to the vertical axis. All other options that depend on axis placement adjust to the rotated axes.

**SIGMA=value | EST**

specifies the parameter $\sigma$, where $\sigma > 0$. Alternatively, you can specify SIGMA=EST to request a maximum likelihood estimate for $\sigma_0$. The interpretation and use of the SIGMA= option depend on the distribution option with which it is used, as summarized in Table 4.26. Enclose this option in parentheses after the distribution option.

---

**Table 4.26** Uses of the SIGMA= Option

<table>
<thead>
<tr>
<th>Distribution Option</th>
<th>Use of the SIGMA= Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA</td>
<td>THETA=$\theta_0$ and SIGMA=$\sigma_0$ request a distribution reference line corresponding to $\theta_0$ and $\sigma_0$.</td>
</tr>
<tr>
<td>EXPONENTIAL</td>
<td></td>
</tr>
<tr>
<td>GAMMA</td>
<td></td>
</tr>
<tr>
<td>PARETO</td>
<td></td>
</tr>
<tr>
<td>POWER</td>
<td></td>
</tr>
<tr>
<td>RAYLEIGH</td>
<td></td>
</tr>
<tr>
<td>WEIBULL</td>
<td></td>
</tr>
<tr>
<td>GUMBEL</td>
<td>MU=$\mu_0$ and SIGMA=$\sigma_0$ request a distribution reference line corresponding to $\mu_0$ and $\sigma_0$.</td>
</tr>
<tr>
<td>LOGNORMAL</td>
<td>SIGMA=$\sigma_1 \ldots \sigma_n$ requests $n$ quantile plots with shape parameters $\sigma_1 \ldots \sigma_n$. The SIGMA= option must be specified.</td>
</tr>
<tr>
<td>NORMAL</td>
<td>MU=$\mu_0$ and SIGMA=$\sigma_0$ request a distribution reference line corresponding to $\mu_0$ and $\sigma_0$. SIGMA=EST requests a line with $\sigma_0$ equal to the sample standard deviation.</td>
</tr>
<tr>
<td>WEIBULL2</td>
<td>SIGMA=$\sigma_0$ and C=c0 request a distribution reference line corresponding to $\sigma_0$ and c0.</td>
</tr>
</tbody>
</table>
**SLOPE=value | EST**
specifies the slope for a distribution reference line requested with the LOGNORMAL and WEIBULL2 options. Enclose the SLOPE= option in parentheses after the distribution option. When you use the SLOPE= lognormal-option with the LOGNORMAL option, you must also specify a threshold parameter value \( \theta_0 \) with the THETA= lognormal-option to request the line. The SLOPE= lognormal-option is an alternative to the ZETA= lognormal-option for specifying \( \zeta_0 \), because the slope is equal to \( \exp(\zeta_0) \).

When you use the SLOPE= Weibull2-option with the WEIBULL2 option, you must also specify a scale parameter value \( \sigma_0 \) with the SIGMA= Weibull2-option to request the line. The SLOPE= Weibull2-option is an alternative to the C= Weibull2-option 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 plots and the third and fourth QQPLOT statements produce the same quantile plots:

```r
proc univariate data=Measures;
  qplot Width / lognormal(sigma=2 theta=0 zeta=0);
  qplot Width / lognormal(sigma=2 theta=0 slope=1);
  qplot Width / weibull2(sigma=2 theta=0 c=.25);
  qplot Width / weibull2(sigma=2 theta=0 slope=4);
```

**SQUARE**
displays the quantile plot in a square frame. By default, the frame is rectangular.

**THETA=value | EST**

**THRESHOLD=value | EST**
specifies the lower threshold parameter \( \theta \) for plots requested with the BETA, EXPONENTIAL, GAMMA, PARETO, POWER, RAYLEIGH, LOGNORMAL, WEIBULL, and WEIBULL2 options. Enclose the THETA= option in parentheses after a distribution option. When used with the WEIBULL2 option, the THETA= option specifies the known lower threshold \( \theta_0 \), for which the default is 0. When used with the other distribution options, the THETA= option specifies \( \theta_0 \) for a distribution reference line; alternatively in this situation, you can specify THETA=EST to request a maximum likelihood estimate for \( \theta_0 \). To request the line, you must also specify a scale parameter.

**WEIBULL(C=value | EST < Weibull-options>)**

**WEIB(C=value | EST < Weibull-options>)**
creates a three-parameter Weibull quantile plot for each value of the required shape parameter \( c \) specified by the mandatory C= Weibull-option. To create a plot that is based on a maximum likelihood estimate for \( c \), specify C=EST. To obtain a graphical estimate of \( c \), specify a list of values in the C= Weibull-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \theta_0 \) and \( \sigma_0 \) with the THETA= and SIGMA= Weibull-options. Alternatively, you can add a line corresponding to estimated values of \( \theta_0 \) and \( \sigma_0 \) with the Weibull-options THETA=EST and SIGMA=EST. Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters \( c , \theta_0 \), and \( \sigma_0 \) is a good fit. You can specify the SCALE= Weibull-option as an alias for the SIGMA= Weibull-option and the THRESHOLD= Weibull-option as an alias for the THETA= Weibull-option. See Example 4.34.
WEIBULL2< (Weibull2-options) >
W2< (Weibull2-options) >
creates a two-parameter Weibull quantile plot. You should use the WEIBULL2 option when your data have a known lower threshold \( \theta_0 \), which is 0 by default. To specify the threshold value \( \theta_0 \), use the THETA= Weibull2-option. By default, THETA=0. An advantage of the two-parameter Weibull plot over the three-parameter Weibull plot is that the parameters \( c \) and \( \sigma \) 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. To obtain a graphical estimate of \( \theta_0 \), specify a list of values for the THETA= Weibull2-option and select the value that most nearly linearizes the point pattern. To assess the point pattern, add a diagonal distribution reference line corresponding to \( \sigma_0 \) and \( c_0 \) with the SIGMA= and C= Weibull2-options. Alternatively, you can add a distribution reference line corresponding to estimated values of \( \sigma_0 \) and \( c_0 \) with the Weibull2-options SIGMA=EST and C=EST. Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters \( c_0 \), \( \theta_0 \), and \( \sigma_0 \) is a good fit. You can specify the SCALE= Weibull2-option as an alias for the SIGMA= Weibull2-option and the SHAPE= Weibull2-option as an alias for the C= Weibull2-option. See Example 4.34.

† WGRID=n
specifies the line thickness for the grid when producing traditional graphics. The option does not apply to ODS Graphics output.

ZETA=value | EST
specifies a value for the scale parameter \( \zeta \) for the lognormal quantile plots requested with the LOGNORMAL option. Enclose the ZETA= lognormal-option in parentheses after the LOGNORMAL option. To request a distribution reference line with intercept \( \theta_0 \) and slope \( \exp(\zeta_0) \), specify the THETA=\( \theta_0 \) and ZETA=\( \zeta_0 \).

VAR Statement
VAR variables ;
The VAR statement specifies the analysis variables and their order in the results. By default, if you omit the VAR statement, PROC UNIVARIATE analyzes all numeric variables that are not listed in the other statements.

Using the Output Statement with the VAR Statement
You must provide 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
WEIGHT variable ;
The WEIGHT statement specifies numeric weights for analysis variables in the statistical calculations. The UNIVARIATE procedure uses the values \( w_i \) of the WEIGHT variable to modify the computation of a number of summary statistics by assuming that the variance of the \( i \)th value \( x_i \) of the analysis variable is equal to \( \sigma^2/w_i \), where \( \sigma \) is an unknown parameter. The values of the WEIGHT variable do not have to be integers and are typically positive. By default, observations with nonpositive or missing values of the WEIGHT variable are handled as follows:

- If the value is zero, the observation is counted in the total number of observations.
- If the value is negative, it is converted to zero, and the observation is counted in the total number of observations.
- If the value is missing, the observation is excluded from the analysis.

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. When you specify a WEIGHT statement, the procedure also computes a weighted standard error and a weighted version of Student’s \( t \) test. The Student’s \( t \) test is the only test of location that PROC UNIVARIATE computes when you weight the analysis variables.

When you specify a WEIGHT variable, the procedure uses its values, \( w_i \), to compute weighted versions of the statistics provided in the Moments table. For example, the procedure computes a weighted mean \( \bar{x}_w \) and a weighted variance \( s^2_w \) as

\[
\bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i}
\]

and

\[
s^2_w = \frac{1}{d} \sum_i w_i (x_i - \bar{x}_w)^2
\]

where \( x_i \) is the \( i \)th variable value. The divisor \( d \) is controlled by the VARDEF= option in the PROC UNIVARIATE statement.

The WEIGHT statement does not affect the determination of the mode, extreme values, extreme observations, or the number of missing values of the analysis variables. However, the weights \( w_i \) are used to compute weighted percentiles. The WEIGHT variable has no effect on graphical displays produced with the plot statements.

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

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

You cannot specify the HISTOGRAM, PROBPLOT, or QQPLOT statements with the WEIGHT statement.

When you use the WEIGHT statement, consider which value of the VARDEF= option is appropriate. See VARDEF= and the calculation of weighted statistics for more information.
Dictionary of Common Options

The following entries provide detailed descriptions of *options* that are common to all the plot statements: CDFPLOT, HISTOGRAM, PPLOT, PROBPLOT, and QQPLOT. Options marked with † are applicable only when traditional graphics are produced.

**ALPHADELTA=value**

specifies the change in successive estimates of $\hat{\alpha}$ at which iteration terminates in the Newton-Raphson approximation of the maximum likelihood estimate of $\alpha$ for gamma distributions requested with the GAMMA option. Enclose the ALPHADELTA= option in parentheses after the GAMMA keyword. Iteration continues until the change in $\alpha$ is less than the value specified or the number of iterations exceeds the value of the MAXITER= option. The default value is 0.00001.

**ALPHAINITIAL=value**

specifies the initial value for $\hat{\alpha}$ in the Newton-Raphson approximation of the maximum likelihood estimate of $\alpha$ for gamma distributions requested with the GAMMA option. Enclose the ALPHAINITIAL= option in parentheses after the GAMMA keyword. The default value is Thom’s approximation of the estimate of $\hat{\alpha}$ (refer to Johnson, Kotz, and Balakrishnan (1995)).

† **ANNOKEY**

applies the annotation requested with the ANNOTATE= option only to the key cell of a comparative plot. By default, the procedure applies annotation to all of the cells. This option is not available unless you use the CLASS statement. You can use the KEYLEVEL= option in the CLASS statement to specify the key cell.

† **ANNOTATE=SAS-data-set**
† **ANNO=SAS-data-set**

specifies an input data set that contains annotate variables as described in *SAS/GRAPH: Reference*. The ANNOTATE= data set you specify in the plot statement is used for all plots created by the statement. You can also specify an ANNOTATE= data set in the PROC UNIVARIATE statement to enhance all plots created by the procedure (see the section “ANNOTATE= Data Sets” on page 429).

† **CAXIS=color**
† **CAXES=color**
† **CA=color**

specifies the color for the axes and tick marks. This option overrides any COLOR= specifications in an AXIS statement.

† **CFRAME=color**

specifies the color for the area that is enclosed by the axes and frame. The area is not filled by default.

† **CFRAMESIDE=color**

specifies the color to fill the frame area for the row labels that display along the left side of a comparative plot. This color also fills the frame area for the label of the corresponding CLASS variable (if you associate a label with the variable). By default, these areas are not filled. This option is not available unless you use the CLASS statement.
† **CFRAMETOP=** *color*

specifies the color to fill the frame area for the column labels that display across the top of a comparative plot. This color also fills the frame area for the label of the corresponding CLASS variable (if you associate a label with the variable). By default, these areas are not filled. This option is not available unless you use the CLASS statement.

† **CHREF=** *color* | *(color-list)*

† **CH=** *color* | *(color-list)*

specifies the colors for horizontal axis reference lines requested by the HREF= option. If you specify a single color, it is used for all HREF= lines. Otherwise, if there are fewer colors specified than reference lines requested, the remaining lines are displayed with the default reference line color. You can also specify the value _default in the color list to request the default color.

† **COLOR=** *color*

† **COLOR=** *color-list*

specifies the color of the curve or reference line associated with a distribution or kernel density estimate. Enclose the COLOR= option in parentheses after a distribution option or the KERNEL option. In a HISTOGRAM statement, you can specify a list of colors in parentheses for multiple density curves.

**CONTENTS='string'**

specifies the table of contents grouping entry for output produced by the plot statement. You can specify CONTENTS='' to suppress the grouping entry.

† **CPROP=** *color* | **EMPTY**

**CROP**

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 plot. By default, no proportion bars are displayed. This option is not available unless you use the CLASS statement. You can specify the keyword EMPTY to display empty bars. See Example 4.20.

For ODS Graphics and traditional graphics with the GSTYLE system option in effect, you can specify CPROP with no argument to produce proportion bars using an appropriate color from the ODS style.

† **CSTATREF=** *color* | *(color-list)*

specifies the colors for reference lines that you request with the STATREF= option. If you specify a single color, it is used for all STATREF= lines. Otherwise, if there are fewer colors specified than reference lines requested, the remaining lines are displayed with the default reference line color. You can also specify the value _default in the color list to request the default color.

† **CTEXT=** *color*

† **CT=** *color*

specifies the color for tick mark values and axis labels. The default is the color specified for the CTEXT= option in the GOPTIONS statement.

† **CTEXTSIDE=** *color*

specifies the color for the row labels that display along the left side of a comparative plot. By default, the color specified by the CTEXT= option is used. If you omit the CTEXT= option, the color specified in the GOPTIONS statement is used. This option is not available unless you use the CLASS statement. You can specify the CFRAMESIDE= option to change the background color for the row labels.
† CTEXTTOP=color
 specifies the color for the column labels that display along the left side of a comparative plot. By default, the color specified by the CTEXT= option is used. If you omit the CTEXT= option, the color specified in the GOPTIONS statement is used. This option is not available unless you specify the CLASS statement. You can use the CFRAMETOP= option to change the background color for the column labels.

† CVREF=color | (color-list)
† CV=color | (color-list)
 specifies the colors for lines requested with the VREF= option. If you specify a single color, it is used for all VREF= lines. Otherwise, if there are fewer colors specified than reference lines requested, the remaining lines are displayed with the default reference line color. You can also specify the value _default in the color list to request the default color.

† DESCRIPTION='string'
† DES='string'
 specifies a description, up to 256 characters long, that appears in the PROC GREPLAY master menu for a traditional graphics chart. The default value is the analysis variable name.

FITINTERVAL=value
 specifies the value of \( z \) for the method of percentiles when this method is used to fit a Johnson \( S_B \) or Johnson \( S_U \) distribution. The FITINTERVAL= option is specified in parentheses after the SB or SU option. The default of \( z \) is 0.524.

FITMETHOD=PERCENTILE | MLE | MOMENTS
 specifies the method used to estimate the parameters of a Johnson \( S_B \) or Johnson \( S_U \) distribution. The FITMETHOD= option is specified in parentheses after the SB or SU option. By default, the method of percentiles is used. You can specify the MLE keyword to request maximum likelihood estimation. The OPTBOUND=, OPTMAXITER=, OPTMAXSTARTS=, OPTPRINT, OPTSEED=, and OPTTOLERANCE= options control the optimizer that performs the maximum likelihood calculation.

FITTOLERANCE=value
 specifies the tolerance value for the ratio criterion when the method of percentiles is used to fit a Johnson \( S_B \) or Johnson \( S_U \) distribution. The FITTOLERANCE= option is specified in parentheses after the SB or SU option. The default value is 0.01.

† FONT=font
 specifies a software font for reference line and axis labels. You can also specify fonts for axis labels in an AXIS statement. The FONT= font takes precedence over the FTEXT= font specified in the GOPTIONS statement.

HAXIS=value
 specifies the name of an AXIS statement describing the horizontal axis.

† 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. Minor tick marks are not labeled. By default, HMINOR=0.

**HREF=values**

draws reference lines that are perpendicular to the horizontal axis at the values that you specify. Also see the CHREF= and LHREF= options.

**HREFLABELS=’label1’ . . . ’labeln’**

**HREFLABEL=’label1’ . . . ’labeln’**

**HREFLAB=’label1’ . . . ’labeln’**

specifies labels for the lines requested by the HREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can have up to 16 characters.

**HREFLABPOS=n**

specifies the vertical position of HREFLABELS= labels, as described in the following table.

<table>
<thead>
<tr>
<th>n</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>along top of plot</td>
</tr>
<tr>
<td>2</td>
<td>staggered from top to bottom of plot</td>
</tr>
<tr>
<td>3</td>
<td>along bottom of plot</td>
</tr>
<tr>
<td>4</td>
<td>staggered from bottom to top of plot</td>
</tr>
</tbody>
</table>

By default, HREFLABPOS=1. **Note:** HREFLABPOS=2 and HREFLABPOS=4 are not supported for ODS Graphics output.

† INFONT=font

specifies a software font to use for text inside the framed areas of the plot. The INFONT= option takes precedence over the FTEXT= option in the GOPTIONS statement. For a list of fonts, see SAS/GRAPH: Reference.

† INHEIGHT=value

specifies the height, in percentage screen units, of text used inside the framed areas of the histogram. By default, the height specified by the HEIGHT= option is used. If you do not specify the HEIGHT= option, the height specified with the HTEXT= option in the GOPTIONS statement is used.

† INTERTILE=value

specifies the distance in horizontal percentage screen units between the framed areas, called tiles, of a comparative plot. By default, INTERTILE=0.75 percentage screen units. This option is not available unless you use the CLASS statement. You can specify INTERTILE=0 to create contiguous tiles.

**ITPRINT**

requests a table that shows the iteration history and optimizer details of the maximum likelihood parameter estimation of a Weibull distribution requested with the WEIBULL or WEIBULL2 option.
† \texttt{L=linetype}

† \texttt{L=linetype-list}

specifies the line type of the curve or reference line associated with a distribution or kernel density estimate. Enclose the \texttt{L=} option in parentheses after the distribution option or the KERNEL option. In a HISTOGRAM statement, you can specify a list of line types in parentheses for multiple density curves.

† \texttt{LHREF=linetype | linetype-list}

† \texttt{LH=linetype | linetype-list}

specifies the line types for the reference lines that you request with the HREF= option. If you specify a single line type, it is used for all HREF= lines. Otherwise, if there are fewer line types specified than reference lines requested, the remaining lines are displayed with the default reference line type. You can also specify line type 0 to request the default line type.

† \texttt{LSTATREF=linetype | linetype-list}

specifies the line types for the reference lines that you request with the STATREF= option. If you specify a single line type, it is used for all STATREF= lines. Otherwise, if there are fewer line types specified than reference lines requested, the remaining lines are displayed with the default reference line type. You can also specify line type 0 to request the default line type.

† \texttt{LVREF=linetype | linetype-list}

† \texttt{LV=linetype | linetype-list}

specifies the line types for lines requested with the VREF= option. If you specify a single line type, it is used for all VREF= lines. Otherwise, if there are fewer line types specified than reference lines requested, the remaining lines are displayed with the default reference line type. You can also specify line type 0 to request the default line type.

\texttt{MAXITER=n}

specifies the maximum number of iterations in the Newton-Raphson approximation of the maximum likelihood estimate of $\alpha$ for gamma distributions requested with the GAMMA option and $c$ for Weibull distributions requested with the WEIBULL and WEIBULL2 options. Enclose the \texttt{MAXITER=} option in parentheses after the GAMMA, WEIBULL, or WEIBULL2 keywords. The default value of $n$ is 20.

† \texttt{NAME='string'}

specifies a name for the plot, up to eight characters long, that appears in the PROC GREPLAY master menu for a traditional graphics chart. The default value is ‘UNIVAR’.

\texttt{NCOLS=n}

\texttt{NCOL=n}

specifies the number of columns per panel in a comparative plot. This option is not available unless you use the CLASS statement. By default, NCOLS=1 if you specify only one CLASS variable, and NCOLS=2 if you specify two CLASS variables. If you specify two CLASS variables, you can use the NCOLS= option with the NROWS= option.

\texttt{NOFRAME}

suppresses the frame around the subplot area.
Chapter 4: The UNIVARIATE Procedure

**NOHLABEL**
suppresses the label for the horizontal axis. You can use this option to reduce clutter.

**NOVLABEL**
suppresses the label for the vertical axis. You can use this option to reduce clutter.

**NOVTICK**
suppresses the tick marks and tick mark labels for the vertical axis. This option also suppresses the label for the vertical axis.

**NROWS=n**

specifies the number of rows per panel in a comparative plot. This option is not available unless you use the CLASS statement. By default, NROWS=2. If you specify two CLASS variables, you can use the NCOLS= option with the NROWS= option.

**ODSFOOTNOTE=FOOTNOTE | FOOTNOTE1 | 'string'**
adds a footnote to ODS Graphics output. If you specify the FOOTNOTE (or FOOTNOTE1) keyword, the value of the SAS FOOTNOTE statement is used as the graph footnote. If you specify a quoted string, that is used as the footnote. The quoted string can contain either of the following escaped characters, which are replaced with the appropriate values from the analysis:

\n is replaced by the analysis variable name.
\l is replaced by the analysis variable label (or name if the analysis variable has no label).

**ODSFOOTNOTE2=FOOTNOTE2 | 'string'**
adds a secondary footnote to ODS Graphics output. If you specify the FOOTNOTE2 keyword, the value of the SAS FOOTNOTE2 statement is used as the secondary graph footnote. If you specify a quoted string, that is used as the secondary footnote. The quoted string can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

\n is replaced by the analysis variable name.
\l is replaced by the analysis variable label (or name if the analysis variable has no label).

**ODSTITLE=TITLE | TITLE1 | NONE | DEFAULT | LABELFMT | 'string'**
specifies a title for ODS Graphics output. You can specify the following values:

**TITLE (or TITLE1)** uses the value of the SAS TITLE statement as the graph title.

**NONE** suppresses all titles from the graph.

**DEFAULT** uses the default ODS Graphics title (a descriptive title that consists of the plot type and the analysis variable name).

**LABELFMT** uses the default ODS Graphics title with the variable label instead of the variable name.

If you specify a quoted string, that is used as the graph title. The quoted string can contain the following escaped characters, which are replaced with the appropriate values from the analysis:
`\n` is replaced by the analysis variable name.

`\l` is replaced by the analysis variable label (or name if the analysis variable has no label).

**ODSTITLE2=TITLE2 | ‘string’**

specifies a secondary title for ODS Graphics output. If you specify the TITLE2 keyword, the value of the SAS TITLE2 statement is used as the secondary graph title. If you specify a quoted string, that is used as the secondary title. The quoted string can contain the following escaped characters, which are replaced with the appropriate values from the analysis:

`\n` is replaced by the analysis variable name.

`\l` is replaced by the analysis variable label (or name if the analysis variable has no label).

**OVERLAY**

specifies that plots associated with different levels of a CLASS variable be overlaid onto a single plot, rather than displayed as separate cells in a comparative plot. If you specify the OVERLAY option with one CLASS variable, the output associated with each level of the CLASS variable is overlaid on a single plot. If you specify the OVERLAY option with two CLASS variables, a comparative plot based on the first CLASS variable’s levels is produced. Each cell in this comparative plot contains overlaid output associated with the levels of the second CLASS variable.

The OVERLAY option applies only to ODS Graphics output.

**SCALE=value**

is an alias for the SIGMA= option for distributions requested by the BETA, EXPONENTIAL, GAMMA, SB, SU, WEIBULL, and WEIBULL2 options and for the ZETA= option for distributions requested by the LOGNORMAL option.

**SHAPE=value**

is an alias for the ALPHA= option for distributions requested by the GAMMA option, for the SIGMA= option for distributions requested by the LOGNORMAL option, and for the C= option for distributions requested by the WEIBULL and WEIBULL2 options.

**STATREF=keyword-list**

draws reference lines at the values of the statistics that are requested in the space-delimited `keyword-list`. These reference lines are perpendicular to the horizontal axis in a histogram or cdf plot, and perpendicular to the vertical axis in a probability or Q-Q plot (unless the ROTATE option is specified). The STATREF= option does not apply to the PPLOT statement.

Valid `keywords` are listed in the following table.
Keyword | Statistic
-------|---------
MAX    | Largest value
MEAN   | Sample mean
MEDIAN | Median (50th percentile)
MIN    | Smallest value
MODE   | Most frequent value
P `pctl` | `pctl`th percentile
Q1     | Lower quartile (25th percentile)
Q3     | Upper quartile (75th percentile)

Note that the factor that is specified with the STD keyword can be positive (which puts a reference line above the mean) or negative (which puts a reference line below the mean).

Also see the CSTATREF=, LSTATREF=, STATREFLABELS=, and STATREFSUBCHAR= options.

STATREFLABELS='label1' ... 'labeln'
STATREFLABEL='label1' ... 'labeln'
STATREFLAB='label1' ... 'labeln'
specifies labels for the lines that you request with the STATREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters long.

STATREFSUBCHAR='character'
specifies a substitution character (such as `#`) for labels that you specify with the STATREFLABELS= option. When the labels are displayed in a graph, the first occurrence of the specified character in each label is replaced with the value of the corresponding STATREF= statistic.

For example, suppose the mean of variable Weight is 155. The following statement creates a histogram with a vertical reference line at 155 with the label “Average=155”:

```
histogram Weight / statref=mean statreftlabel='Average=#' statreftsubchar='#';
```

† TURNV LABELS
† TURNV LABEL
    turns the characters in the vertical axis labels so that they display vertically. This happens by default when you use a hardware font.

VAXIS= `name`
VAXIS= `value-list`
specifies the name of an AXIS statement describing the vertical axis. In a HISTOGRAM statement, you can alternatively specify a `value-list` for the vertical axis.

VAXISLABEL= 'label'
specifies a label for the vertical axis. Labels can have up to 40 characters.
† VMINOR=n
† VM=n
 specifies the number of minor tick marks between each major tick mark on the vertical axis. Minor
tick marks are not labeled. The default is zero.

VREF=value-list
 draws reference lines perpendicular to the vertical axis at the values specified. Also see the CVREF= and LVREF= options.

VREFLABELS='label1'...'labeln'
 VREFLABEL='label1'...'labeln'
 specifies labels for the lines requested by the VREF= option. The number of labels must equal the
number of lines. Enclose each label in quotes. Labels can have up to 16 characters.

VREFLABPOS=n
 specifies the horizontal position of VREFLABELS= labels. If you specify VREFLABPOS=1, the
labels are positioned at the left of the plot. If you specify VREFLABPOS=2, the labels are positioned
at the right of the plot. By default, VREFLABPOS=1.

† W=value
† W=value-list
 specifies the width in pixels of the curve or reference line associated with a distribution or kernel
density estimate. Enclose the W= option in parentheses after the distribution option or the KERNEL
option. In a HISTOGRAM statement, you can specify a list of widths in parentheses for multiple
density curves.

† WAXIS=n
 specifies the line thickness, in pixels, for the axes and frame.

---

Details: UNIVARIATE Procedure

Missing Values

PROC UNIVARIATE excludes missing values for an 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 missing values and reports this information in the ODS table named “Missing Values.” See the section “ODS Table Names” on page 436. Before the number of missing values is tabulated, PROC UNIVARIATE excludes observations when either of the following conditions exist:

- 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)

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 of the form \(ui\), where \(u\) is the nonnegative rounding unit and \(i\) is an integer. The interval width is \(u\). Any variable value that falls in an interval is rounded 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 \((0, \pm 2, \pm 4, \ldots)\).

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

<table>
<thead>
<tr>
<th>(i)</th>
<th>Interval</th>
<th>Midpoint</th>
<th>Left endpt rounds to</th>
<th>Right endpt rounds to</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>([-2.5, -1.5])</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>-1</td>
<td>([-1.5, -0.5])</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>([-0.5, 0.5])</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>([0.5, 1.5])</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>([1.5, 2.5])</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

When ROUND=0.5 and the analysis variable values are between \(-1.25\) and 1.25, the intervals are as follows:

<table>
<thead>
<tr>
<th>(i)</th>
<th>Interval</th>
<th>Midpoint</th>
<th>Left endpt rounds to</th>
<th>Right endpt rounds to</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>([-1.25, -0.75])</td>
<td>-1.0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>([-0.75, -0.25])</td>
<td>-0.5</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>([-0.25, 0.25])</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>([0.25, 0.75])</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>([0.75, 1.25])</td>
<td>1.0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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.
Descriptive Statistics

This section provides computational details for the descriptive statistics that are computed with the PROC UNIVARIATE statement. These statistics can also be saved in an OUT= data set by specifying keywords listed in Table 4.14 in the OUTPUT statement.

Standard algorithms (Fisher 1973) are used to compute the moment statistics. The computational methods used by the UNIVARIATE procedure are consistent with those used by other SAS procedures for calculating descriptive statistics.

The following sections give specific details on a number of statistics calculated by the UNIVARIATE procedure.

Mean

The sample mean is calculated as

\[ \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \]

where \( n \) is the number of nonmissing values for a variable, \( x_i \) is the \( i \)th value of the variable, and \( w_i \) is the weight associated with the \( i \)th value of the variable. If there is no WEIGHT variable, the formula reduces to

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

Sum

The sum is calculated as \( \sum_{i=1}^{n} w_i x_i \), where \( n \) is the number of nonmissing values for a variable, \( x_i \) is the \( i \)th value of the variable, and \( w_i \) is the weight associated with the \( i \)th value of the variable. If there is no WEIGHT variable, the formula reduces to \( \sum_{i=1}^{n} x_i \).

Sum of the Weights

The sum of the weights is calculated as \( \sum_{i=1}^{n} w_i \), where \( n \) is the number of nonmissing values for a variable \( w_i \) is the weight associated with the \( i \)th value of the variable. If there is no WEIGHT variable, the sum of the weights is \( n \).

Variance

The variance is calculated as

\[ \frac{1}{d} \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 \]

where \( n \) is the number of nonmissing values for a variable, \( x_i \) is the \( i \)th value of the variable, \( \bar{x}_w \) is the weighted mean, \( w_i \) is the weight associated with the \( i \)th value of the variable, and \( d \) is the divisor controlled
by the VARDEF= option in the PROC UNIVARIATE statement:

\[
d = \begin{cases} 
  n - 1 & \text{if VARDEF=DF (default)} \\
  n & \text{if VARDEF=N} \\
  \left( \sum_i w_i \right) - 1 & \text{if VARDEF=WDF} \\
  \sum_i w_i & \text{if VARDEF=WEIGHT | WGT}
\end{cases}
\]

If there is no WEIGHT variable, the formula reduces to

\[
\frac{1}{d} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

**Standard Deviation**

The standard deviation is calculated as

\[
s_w = \sqrt{\frac{1}{d} \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}
\]

where \( n \) is the number of nonmissing values for a variable, \( x_i \) is the \( i \)th value of the variable, \( \bar{x}_w \) is the weighted mean, \( w_i \) is the weight associated with the \( i \)th value of the variable, and \( d \) is the divisor controlled by the VARDEF= option in the PROC UNIVARIATE statement. If there is no WEIGHT variable, the formula reduces to

\[
s = \sqrt{\frac{1}{d} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

**Skewness**

The sample skewness, which measures the tendency of the deviations to be larger in one direction than in the other, is calculated as follows depending on the VARDEF= option:

<table>
<thead>
<tr>
<th>VARDEF</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF (default)</td>
<td>( \frac{n}{(n - 1)(n - 2)} \sum_{i=1}^{n} w_i^{3/2} \left( \frac{x_i - \bar{x}_w}{s_w} \right)^3 )</td>
</tr>
<tr>
<td>N</td>
<td>( \frac{1}{n} \sum_{i=1}^{n} w_i^{3/2} \left( \frac{x_i - \bar{x}_w}{s_w} \right)^3 )</td>
</tr>
<tr>
<td>WDF</td>
<td>missing</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>missing</td>
</tr>
</tbody>
</table>
where \( n \) is the number of nonmissing values for a variable, \( x_i \) is the \( i \)th value of the variable, \( \bar{x}_w \) is the sample average, \( s_w \) is the sample standard deviation, and \( w_i \) is the weight associated with the \( i \)th value of the variable. If \( \text{VARDEF=} \text{DF} \), then \( n \) must be greater than 2. If there is no \( \text{WEIGHT} \) variable, then \( w_i = 1 \) for all \( i = 1, \ldots, n \).

The sample skewness can be positive or negative; it measures the asymmetry of the data distribution and estimates the theoretical skewness \( \sqrt{\beta_1} = \mu_3 \mu_2^{-\frac{3}{2}} \), where \( \mu_2 \) and \( \mu_3 \) are the second and third central moments. Observations that are normally distributed should have a skewness near zero.

### Kurtosis

The sample kurtosis, which measures the heaviness of tails, is calculated as follows depending on the \( \text{VARDEF=} \) option:

<table>
<thead>
<tr>
<th>( \text{VARDEF} )</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF (default)</td>
<td>( \frac{n(n + 1)}{(n - 1)(n - 2)(n - 3)} \sum_{i=1}^{n} w_i^2 \left( \frac{x_i - \bar{x}_w}{s_w} \right)^4 - \frac{3(n - 1)^2}{(n - 2)(n - 3)} )</td>
</tr>
<tr>
<td>N</td>
<td>( \frac{1}{n} \sum_{i=1}^{n} w_i^2 \left( \frac{x_i - \bar{x}_w}{s_w} \right)^4 - 3 )</td>
</tr>
<tr>
<td>WDF</td>
<td>missing</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>WGT</td>
</tr>
</tbody>
</table>

where \( n \) is the number of nonmissing values for a variable, \( x_i \) is the \( i \)th value of the variable, \( \bar{x}_w \) is the sample average, \( s_w \) is the sample standard deviation, and \( w_i \) is the weight associated with the \( i \)th value of the variable. If \( \text{VARDEF=} \text{DF} \), then \( n \) must be greater than 3. If there is no \( \text{WEIGHT} \) variable, then \( w_i = 1 \) for all \( i = 1, \ldots, n \).

The sample kurtosis measures the heaviness of the tails of the data distribution. It estimates the adjusted theoretical kurtosis denoted as \( \beta_2 - 3 \), where \( \beta_2 = \frac{\mu_4}{\mu_2^2} \), and \( \mu_4 \) is the fourth central moment. Observations that are normally distributed should have a kurtosis near zero.

### Coefficient of Variation (CV)

The coefficient of variation is calculated as

\[
CV = \frac{100 \times s_w}{\bar{x}_w}
\]
Calculating the Mode

The mode is the value that occurs most often in the data. PROC UNIVARIATE counts repetitions of the values of the analysis variables or, if you specify the ROUND= option, the rounded values. If a tie occurs for the most frequent value, the procedure reports the lowest mode in the table labeled “Basic Statistical Measures” in the statistical output. 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. See Example 4.2.

Calculating Percentiles

The UNIVARIATE procedure automatically computes the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles (quantiles), as well as the minimum and maximum of each analysis variable. To compute percentiles other than these default percentiles, use the PCTLPTS= and PCTLPRE= options in the OUTPUT statement.

You can specify one of five definitions for computing the percentiles with the PCTLDEF= option. Let $n$ be the number of nonmissing values for a variable, and let $x_1, x_2, \ldots, x_n$ represent the ordered values of the variable. Let the $t$th percentile be $y$, set $p = \frac{t}{100}$, and let

\[
np = j + g \quad \text{when PCTLDEF=1, 2, 3, or 5}
\]

\[
(n + 1)p = j + g \quad \text{when PCTLDEF=4}
\]

where $j$ is the integer part of $np$, and $g$ is the fractional part of $np$. Then the PCTLDEF= option defines the $t$th percentile, $y$, as described in the following table.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
PCTLDEF & Description & Formula \\
\hline
1 & weighted average at $x_{np}$ & $y = (1 - g)x_j + gx_{j+1}$ where $x_0$ is taken to be $x_1$ \\
2 & observation numbered closest to $np$ & $y = x_j$ if $g < \frac{1}{2}$ \\
 & & $y = x_j$ if $g = \frac{1}{2}$ and $j$ is even \\
 & & $y = x_{j+1}$ if $g = \frac{1}{2}$ and $j$ is odd \\
 & & $y = x_{j+1}$ if $g > \frac{1}{2}$ \\
3 & empirical distribution function & $y = x_j$ if $g = 0$ \\
 & & $y = x_{j+1}$ if $g > 0$ \\
4 & weighted average aimed at $x_{(n+1)p}$ & $y = (1 - g)x_j + gx_{j+1}$ where $x_{n+1}$ is taken to be $x_n$ \\
5 & empirical distribution function with averaging & $y = \frac{1}{2}(x_j + x_{j+1})$ if $g = 0$ \\
 & & $y = x_{j+1}$ if $g > 0$ \\
\hline
\end{tabular}
\caption{Percentile Definitions}
\end{table}
**Weighted Percentiles**

When you use a WEIGHT statement, the percentiles are computed differently. The 100$p$th weighted percentile $y$ is computed from the empirical distribution function with averaging:

$$y = \begin{cases} 
  x_1 & \text{if } w_1 > pW \\
  \frac{1}{2}(x_i + x_{i+1}) & \text{if } \sum_{j=1}^{i} w_j = pW \\
  x_{i+1} & \text{if } \sum_{j=1}^{i} w_j < pW < \sum_{j=1}^{i+1} w_j
\end{cases}$$

where $w_i$ is the weight associated with $x_i$ and $W = \sum_{i=1}^{n} w_i$ is the sum of the weights.

Note that the PCTLDEF= option is not applicable when a WEIGHT statement is used. However, in this case, if all the weights are identical, the weighted percentiles are the same as the percentiles that would be computed without a WEIGHT statement and with PCTLDEF=5.

**Confidence Limits for Percentiles**

You can use the CIPCTLNORMAL option to request confidence limits for percentiles, assuming the data are normally distributed. These limits are described in Section 4.4.1 of Hahn and Meeker (1991). When $0 < p < \frac{1}{2}$, the two-sided 100($1 - \alpha$)% confidence limits for the 100$p$th percentile are

lower limit $= \bar{X} - g'(\frac{\alpha}{2}; 1 - p, n)s$
upper limit $= \bar{X} - g'(\frac{\alpha}{2}; p, n)s$

where $n$ is the sample size. When $\frac{1}{2} \leq p < 1$, the two-sided 100($1 - \alpha$)% confidence limits for the 100$p$th percentile are

lower limit $= \bar{X} + g'(\frac{\alpha}{2}; 1 - p, n)s$
upper limit $= \bar{X} + g'(\frac{\alpha}{2}; p, n)s$

One-sided 100($1 - \alpha$)% confidence bounds are computed by replacing $\frac{\alpha}{2}$ by $\alpha$ in the appropriate preceding equation. The factor $g'(\gamma, p, n)$ is related to the noncentral $t$ distribution and is described in Owen and Hua (1977) and Odeh and Owen (1980). See Example 4.10.

You can use the CIPCTLDF option to request distribution-free confidence limits for percentiles. In particular, it is not necessary to assume that the data are normally distributed. These limits are described in Section 5.2 of Hahn and Meeker (1991). The two-sided 100($1 - \alpha$)% confidence limits for the 100$p$th percentile are

lower limit $= X(l)$
upper limit $= X(u)$

where $X(j)$ is the $j$th order statistic when the data values are arranged in increasing order:

$X(1) \leq X(2) \leq \ldots \leq X(n)$

The lower rank $l$ and upper rank $u$ are integers that are symmetric (or nearly symmetric) around $\lfloor np \rfloor + 1$, where $\lfloor np \rfloor$ is the integer part of $np$ and $n$ is the sample size. Furthermore, $l$ and $u$ are chosen so that $X(l)$ and $X(u)$ are as close to $X_{\lfloor np \rfloor + 1}$ as possible while satisfying the coverage probability requirement,

$$Q(u - 1; n, p) - Q(l - 1; n, p) \geq 1 - \alpha$$
where \( Q(k; n, p) \) is the cumulative binomial probability,

\[
Q(k; n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1 - p)^{n-i}
\]

In some cases, the coverage requirement cannot be met, particularly when \( n \) is small and \( p \) is near 0 or 1. To relax the requirement of symmetry, you can specify CIPCTLDF(TYPE = ASYMMETRIC). This option requests symmetric limits when the coverage requirement can be met, and asymmetric limits otherwise.

If you specify CIPCTLDF(TYPE = LOWER), a one-sided 100(1 - \( \alpha \))% lower confidence bound is computed as \( X_{(l)} \), where \( l \) is the largest integer that satisfies the inequality

\[
1 - Q(l - 1; n, p) \geq 1 - \alpha
\]

with \( 0 < l \leq n \). Likewise, if you specify CIPCTLDF(TYPE = UPPER), a one-sided 100(1 - \( \alpha \))% lower confidence bound is computed as \( X_{(u)} \), where \( u \) is the largest integer that satisfies the inequality

\[
Q(u - 1; n, p) \geq 1 - \alpha \quad \text{where} \quad 0 < u \leq n
\]

Note that confidence limits for percentiles are not computed when a WEIGHT statement is specified. See Example 4.10.

---

### Tests for Location

PROC UNIVARIATE provides three tests for location: 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 \( \mu_0 \) against the two-sided alternative that the mean or median is not equal to \( \mu_0 \). By default, PROC UNIVARIATE sets the value of \( \mu_0 \) to zero. You can use the MU0= option in the PROC UNIVARIATE statement to specify the value of \( \mu_0 \). 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 (VARDEF=DF). See Example 4.12.

You can also use these tests to 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. Note that you can also carry out these tests by using the PAIRED statement in the TTEST procedure; see Chapter 103, “The TTEST Procedure” (SAS/STAT User’s Guide). Also see Example 4.13.
Student’s t Test

PROC UNIVARIATE calculates the $t$ statistic as

$$
t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
$$

where $\bar{x}$ is the sample mean, $n$ is the number of nonmissing values for a variable, and $s$ is the sample standard deviation. The null hypothesis is that the population mean equals $\mu_0$. 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

$$
t_w = \frac{\bar{x}_w - \mu_0}{s_w / \sqrt{\sum_{i=1}^{n} w_i}}
$$

where $\bar{x}_w$ is the weighted mean, $s_w$ is the weighted standard deviation, and $w_i$ is the weight for $i$th 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

$$
M = (n^+ - n^-)/2
$$

where $n^+$ is the number of values that are greater than $\mu_0$, and $n^-$ is the number of values that are less than $\mu_0$. Values equal to $\mu_0$ are discarded. Under the null hypothesis that the population median is equal to $\mu_0$, the $p$-value for the observed statistic $M_{obs}$ is

$$
Pr(|M_{obs}| \geq |M|) = 0.5^{(n_t - 1)} \sum_{j=0}^{\min(n^+, n^-)} \binom{n_t}{j} \binom{n_t}{j}
$$

where $n_t = n^+ + n^-$ is the number of $x_i$ values not equal to $\mu_0$.

**NOTE:** If $n^+$ and $n^-$ are equal, the $p$-value is equal to one.

Wilcoxon Signed Rank Test

The signed rank statistic $S$ is computed as

$$
S = \sum_{i : x_i > \mu_0} r_i^+ - \frac{n_t(n_t + 1)}{4}
$$

where $r_i^+$ is the rank of $|x_i - \mu_0|$ after discarding values of $x_i = \mu_0$, and $n_t$ is the number of $x_i$ values not equal to $\mu_0$. Average ranks are used for tied values.
If \( n_t \leq 20 \), the significance of \( S \) is computed from the exact distribution of \( S \), where the distribution is a convolution of scaled binomial distributions. When \( n_t > 20 \), the significance of \( S \) is computed by treating \( S \) as a Student’s \( t \) variate with \( n_t - 1 \) degrees of freedom. \( V \) is computed as

\[
V = \frac{1}{24} n_t (n_t + 1)(2n_t + 1) - \frac{1}{48} \sum t_i (t_i + 1)(t_i - 1)
\]

where the sum is over groups tied in absolute value and where \( t_i \) is the number of values in the \( i \)th group (Iman 1974; Conover 1980). The null hypothesis tested is that the mean (or median) is \( \mu_0 \), assuming that the distribution is symmetric. Refer to Lehmann and D’Aabra (1975).

**Confidence Limits for Parameters of the Normal Distribution**

The two-sided 100(1 - \( \alpha \))% confidence interval for the mean has upper and lower limits

\[
\bar{x} \pm t_{1-\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}}
\]

where \( s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \) and \( t_{1-\frac{\alpha}{2};n-1} \) is the (1 - \( \frac{\alpha}{2} \)) percentile of the \( t \) distribution with \( n - 1 \) degrees of freedom. The one-sided upper 100(1 - \( \alpha \))% confidence limit is computed as \( \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha;n-1} \) and the one-sided lower 100(1 - \( \alpha \))% confidence limit is computed as \( \bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha;n-1} \). See Example 4.9.

The two-sided 100(1 - \( \alpha \))% confidence interval for the standard deviation has lower and upper limits,

\[
s \sqrt{\chi^2_{1-\frac{\alpha}{2};n-1}} \quad \text{and} \quad s \sqrt{\chi^2_{\frac{\alpha}{2};n-1}}
\]

respectively, where \( \chi^2_{1-\frac{\alpha}{2};n-1} \) and \( \chi^2_{\frac{\alpha}{2};n-1} \) are the (1 - \( \frac{\alpha}{2} \)) and \( \frac{\alpha}{2} \) percentiles of the chi-square distribution with \( n - 1 \) degrees of freedom. A one-sided 100(1 - \( \alpha \))% confidence limit has lower and upper limits,

\[
s \sqrt{\chi^2_{1-\alpha;n-1}} \quad \text{and} \quad s \sqrt{\chi^2_{\alpha;n-1}}
\]

respectively. The 100(1 - \( \alpha \))% 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 - \( \alpha \))% confidence interval for the weighted mean is

\[
\bar{x}_w \pm t_{1-\frac{\alpha}{2}} \frac{s_w}{\sqrt{\sum_{i=1}^{n} w_i}}
\]

where \( \bar{x}_w \) is the weighted mean, \( s_w \) is the weighted standard deviation, \( w_i \) is the weight for \( i \)th observation, and \( t_{1-\frac{\alpha}{2}} \) is the (1 - \( \frac{\alpha}{2} \)) percentile for the \( t \) distribution with \( n - 1 \) degrees of freedom.

Confidence intervals for the weighted standard deviation are computed by substituting \( s_w \) for \( s \) in the preceding formulas for confidence limits for the standard deviation.
Robust Estimators

A statistical method is robust if it is insensitive to moderate or even large departures from the assumptions that justify the method. PROC UNIVARIATE provides several methods for robust estimation of location and scale. See Example 4.11.

Winsorized Means

The Winsorized mean is a robust estimator of the location that is relatively insensitive to outliers. The \( k \)-times Winsorized mean is calculated as

\[
\bar{x}_{wk} = \frac{1}{n} \left( (k + 1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x(i) + (k + 1)x_{(n-k)} \right)
\]

where \( n \) is the number of observations and \( x(i) \) is the \( i \)th order statistic when the observations are arranged in increasing order:

\[
x(1) \leq x(2) \leq \ldots \leq x(n)
\]

The Winsorized mean is computed as the ordinary mean after the \( k \) smallest observations are replaced by the \( (k + 1) \)st smallest observation and the \( k \) largest observations are replaced by the \( (k + 1) \)st largest observation.

For data from a symmetric distribution, the Winsorized mean is an unbiased estimate of the population mean. However, 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

\[
\hat{s}^2_{wk} = (k + 1)(x_{(k+1)} - \bar{x}_{wk})^2 + \sum_{i=k+2}^{n-k-1} (x(i) - \bar{x}_{wk})^2 + (k + 1)(x_{(n-k)} - \bar{x}_{wk})^2
\]

The Winsorized \( t \) statistic is given by

\[
t_{wk} = \frac{\bar{x}_{wk} - \mu_0}{\text{SE}(\bar{x}_{wk})}
\]

where \( \mu_0 \) denotes the location under the null hypothesis and the standard error of the Winsorized mean is

\[
\text{SE}(\bar{x}_{wk}) = \frac{n - 1}{n - 2k - 1} \times \frac{s_{wk}}{\sqrt{n(n - 1)}}
\]

When the data are from a symmetric distribution, the distribution of \( t_{wk} \) is approximated by a Student’s \( t \) distribution with \( n - 2k - 1 \) degrees of freedom (Tukey and McLaughlin 1963; Dixon and Tukey 1968).

The Winsorized \( 100(1 - \frac{\alpha}{2})\% \) confidence interval for the location parameter has upper and lower limits

\[
\bar{x}_{wk} \pm t_{1 - \frac{\alpha}{2} ; n - 2k - 1} \times \text{SE}(\bar{x}_{wk})
\]

where \( t_{1 - \frac{\alpha}{2} ; n - 2k - 1} \) is the \( (1 - \frac{\alpha}{2}) \)th percentile of the Student’s \( t \) distribution with \( n - 2k - 1 \) degrees of freedom.
Trimmed Means

Like the Winsorized mean, the trimmed mean is a robust estimator of the location that is relatively insensitive to outliers. The $k$-times trimmed mean is calculated as

$$\bar{x}_{tk} = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x^{(i)}$$

where $n$ is the number of observations and $x^{(i)}$ is the $i$th order statistic when the observations are arranged in increasing order:

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}$$

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. However, 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 $s_{wk}^2$, which is defined in the section “Winsorized Means” on page 387; see Tukey and McLaughlin (1963). This can be used to compute a trimmed $t$ test which is based on the test statistic

$$t_{tk} = \frac{\bar{x}_{tk} - \mu_0}{SE(\bar{x}_{tk})}$$

where the standard error of the trimmed mean is

$$SE(\bar{x}_{tk}) = \frac{s_{wk}}{\sqrt{(n - 2k)(n - 2k - 1)}}$$

When the data are from a symmetric distribution, the distribution of $t_{tk}$ is approximated by a Student’s $t$ distribution with $n - 2k - 1$ degrees of freedom (Tukey and McLaughlin 1963; Dixon and Tukey 1968).

The “trimmed” $100(1 - \alpha)$% confidence interval for the location parameter has upper and lower limits

$$\bar{x}_{tk} \pm t_{1-\frac{\alpha}{2};n-2k-1} SE(\bar{x}_{tk})$$

where $t_{1-\frac{\alpha}{2};n-2k-1}$ is the $(1 - \frac{\alpha}{2})100th$ percentile of the Student’s $t$ distribution with $n - 2k - 1$ degrees of freedom.

Robust Estimates of Scale

The sample standard deviation, which is the most commonly used estimator of scale, is sensitive to outliers. Robust scale estimators, on the other hand, remain bounded when a single data value is replaced by an arbitrarily large or small value. The UNIVARIATE procedure computes several robust measures of scale, including the interquartile range, Gini’s mean difference $G$, the median absolute deviation about the median (MAD), $Q_n$, and $S_n$. In addition, the procedure computes estimates of the normal standard deviation $\sigma$ derived from each of these measures.
The interquartile range (IQR) is simply the difference between the upper and lower quartiles. For a normal population, $\sigma$ can be estimated as $\text{IQR}/1.34898$.

Gini’s mean difference is computed as

$$G = \frac{1}{\binom{n}{2}} \sum_{i < j} |x_i - x_j|$$

For a normal population, the expected value of $G$ is $2\sigma/\sqrt{n}$. Thus $G \sqrt{n}/2$ is a robust estimator of $\sigma$ when the data are from a normal sample. For the normal distribution, this estimator has high efficiency relative to the usual sample standard deviation, and it is also less sensitive to the presence of outliers.

A very robust scale estimator is the MAD, the median absolute deviation from the median (Hampel 1974), which is computed as

$$\text{MAD} = \text{med}_i (|x_i - \text{med}_j(x_j)|)$$

where the inner median, $\text{med}_j(x_j)$, is the median of the $n$ observations, and the outer median (taken over $i$) is the median of the $n$ absolute values of the deviations about the inner median. For a normal population, $1.4826 \times \text{MAD}$ is an estimator of $\sigma$.

The MAD has low efficiency for normal distributions, and it may not always be appropriate for symmetric distributions. Rousseeuw and Croux (1993) proposed two statistics as alternatives to the MAD. The first is

$$S_n = 1.1926 \times \text{med}_i (\text{med}_j(|x_i - x_j|))$$

where the outer median (taken over $i$) is the median of the $n$ medians of $|x_i - x_j|$, $j = 1, 2, \ldots, n$. To reduce small-sample bias, $c_{sn}S_n$ is used to estimate $\sigma$, where $c_{sn}$ is a correction factor; see Croux and Rousseeuw (1992).

The second statistic proposed by Rousseeuw and Croux (1993) is

$$Q_n = 2.2219 \{|x_i - x_j|; i < j\} (k)$$

where

$$k = \left( \left\lfloor \frac{n}{2} \right\rfloor + 1 \right)$$

In other words, $Q_n$ is 2.2219 times the $k$th order statistic of the $\binom{n}{2}$ distances between the data points. The bias-corrected statistic $c_{qn}Q_n$ is used to estimate $\sigma$, where $c_{qn}$ is a correction factor; see Croux and Rousseeuw (1992).

---

**Creating Line Printer Plots**

When ODS Graphics is disabled, 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.

Note that these plots are a legacy feature of the UNIVARIATE procedure in earlier versions of SAS. They
predate the addition of the CDFPLOT, HISTOGRAM, PPLOT, PROBPLOT, and QQPLOT statements,
which provide high-resolution graphics displays. Also note that line printer plots requested with the PLOTS
option are mainly intended for use with the ODS LISTING destination. See Example 4.5.

**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 box 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.

**NOTE:** To produce box plots that use high-resolution graphics, use the BOXPLOT procedure in SAS/STAT

**Normal Probability Plot**

The normal probability plot 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

\[
v_i = \frac{r_i - \frac{3}{8}}{n + \frac{1}{4}}
\]

\( \Phi^{-1}(\cdot) \) = inverse of the standard normal distribution function

\( r_i \) = rank of the \( i \)th data value when ordered from smallest to largest

\( n \) = number of nonmissing observations

For a weighted normal probability plot, the \( i \)th ordered observation is plotted against \( \Phi^{-1}(v_i) \) where

\[
v_i = \frac{(1 - \frac{1}{n}) \sum_{j=1}^{i} w(j)}{(1 + \frac{1}{n}) \sum_{j=1}^{n} w(j)}
\]

\( w(j) \) = weight associated with the \( j \)th ordered observation

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 \( \text{VARDEF=} \) is WDF or WEIGHT, a reference line with intercept \( \hat{\mu} \) and slope \( \hat{\sigma} \) is added to the plot. When the value of \( \text{VARDEF=} \) is DF or N, the slope is \( \frac{\text{sigma}}{\sqrt{w}} \) where \( w = \frac{\sum_{i=1}^{n} w_i}{n} \) is the average weight.

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

If the data are normally distributed with mean \( \mu \) and standard deviation \( \sigma \), and each observation has an identical weight \( w \), then the points on the plot should lie approximately on a straight line. The intercept for this line is \( \mu \). The slope is \( \sigma \) when \( \text{VARDEF=} \) is WDF or WEIGHT, and the slope is \( \frac{\sigma}{\sqrt{w}} \) when \( \text{VARDEF=} \) is DF or N.

**NOTE:** To produce high-resolution probability plots, use the PROBPLOT statement in PROC UNIVARIATE; see the section “PROBPLOT Statement” on page 345.

### Side-by-Side Box Plots

When you use a BY statement with the PLOT option, PROC UNIVARIATE produces side-by-side box plots, one for each BY group. The box plots (also known as schematic plots) use a common scale that enables 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.

**NOTE:** To produce high-resolution side-by-side box plots, use the BOXPLOT procedure in SAS/STAT software. See Chapter 28, “The BOXPLOT Procedure” (SAS/STAT User’s Guide).

---

**Creating High-Resolution Graphics**

You can use the CDFPLOT, HISTOGRAM, PPLOT, PROBPLOT, and QQPLOT statements to create high-resolution graphs.
The CDFPLOT statement plots the observed cumulative distribution function of a variable. You can optionally superimpose a fitted theoretical distribution on the plot.

The HISTOGRAM statement creates histograms that enable 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 the sections “Formulas for Fitted Continuous Distributions” on page 401 and “Kernel Density Estimates” on page 418.

The PPPlot statement creates a probability-probability (P-P) plot, which compares the empirical cumulative distribution function (ECDF) of a variable with a specified theoretical cumulative distribution function. You can use a P-P plot to determine how well a theoretical distribution models a set of measurements.

The PROBPLOT statement creates a probability plot, which compares ordered values of a variable with percentiles of a specified theoretical distribution. Probability plots are useful for graphical estimation of percentiles.

The QQPLOT statement creates a quantile-quantile plot, which compares ordered values of a variable with quantiles of a specified theoretical distribution. Q-Q plots are useful for graphical estimation of distribution parameters.

**Note:** You can use the CLASS statement with any of these plot statements to produce comparative versions of the plots.

**Alternatives for Producing Graphics**

The UNIVARIATE procedure supports two kinds of graphical output.

- ODS Statistical Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.

- Otherwise, traditional graphics are produced if SAS/GRAPH® is licensed.

Traditional graphics are saved in graphics catalogs. Their appearance is controlled by the SAS/GRAPH GOPTIONS, AXIS, and SYMBOL statements (as described in SAS/GRAPH: Reference) and numerous specialized plot statement options.

ODS Statistical Graphics (or ODS Graphics for short) is an extension to the Output Delivery System (ODS) that can be enabled by specifying the ODS GRAPHICS statement prior to your procedure statements. An ODS graph is produced in ODS output (not a graphics catalog), and the details of its appearance and layout are controlled by ODS styles and templates rather than by SAS/GRAPH statements and procedure options. See Chapter 21, “Statistical Graphics Using ODS” (SAS/STAT User’s Guide), for a thorough discussion of ODS Graphics.
Prior to SAS 9.2, the plots produced by PROC UNIVARIATE were extremely basic by default. Producing attractive graphical output required the careful selection of colors, fonts, and other elements, which were specified via SAS/GRAPH statements and plot statement options. Beginning with SAS 9.2, the default appearance of traditional graphs is governed by the prevailing ODS style, which automatically produces attractive, consistent output. The SAS/GRAPH statements and procedure options for controlling graph appearance continue to be honored for traditional graphics. You can specify the NOGSTYLE system option to prevent the ODS style from affecting the appearance of traditional graphs. This enables existing PROC UNIVARIATE programs to produce customized graphs that appear as they did under previous SAS releases.

The appearance of ODS Graphics output is also controlled by the ODS style, but it is not affected by SAS/GRAPH statements or plot statement options that govern traditional graphics. For example, the CAXIS= option used to specify the color of graph axes in traditional graphics is ignored when producing ODS Graphics output. **NOTE:** Some features available with traditional graphics are not supported in ODS Graphics by the UNIVARIATE procedure for SAS 9.2.

The traditional graphics system enables you to control every detail of a graph through convenient procedure syntax. ODS Graphics provides the highest quality output with minimal syntax and full compatibility with graphics produced by SAS/STAT and SAS/ETS procedures.

The following code produces a histogram with a fitted lognormal distribution of the LoanToValueRatio data introduced in the section “Summarizing a Data Distribution” on page 267:

```sas
options nogstyle;
ods graphics off;
proc univariate data=HomeLoans noprint;
  histogram LoanToValueRatio / lognormal;
  inset lognormal(theta sigma zeta) / position=ne;
run;
```

The NOGSTYLE system option keeps the ODS style from influencing the output, and no SAS/GRAPH statements or procedure options affecting the appearance of the plot are specified. **Figure 4.8** shows the resulting histogram, which is essentially identical to the default output produced under releases prior to SAS 9.2.
Figure 4.8 Traditional Graph with NOGSTYLE

Figure 4.9 shows the result of executing the same code with the GSTYLE system option turned on (the default). Note the influence of the ODS style on the histogram’s appearance. For example, the quality of the text is improved and histogram bars are filled by default.
Figure 4.9 Traditional Graph with GSTYLE

Figure 4.10 shows the same histogram produced using ODS Graphics. The histogram’s appearance is governed by the same style elements as in Figure 4.9, but the plots are not identical. Note, for example, the title incorporated in the ODS Graphics output and the smoother appearance of the fitted curve.
Figure 4.10 ODS Graphics Output

Using the CLASS Statement to Create Comparative Plots

When you use the CLASS statement with the CDFPLOT, HISTOGRAM, PPLOT, PROBPLOT, or QQPLOT statements, PROC UNIVARIATE creates comparative versions of the plots. You can use these plot statements 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.

When you create a one-way comparative plot, the observations in the input data set are sorted by the method specified in the ORDER= option. 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. See Example 4.15.

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.
Positioning Insets

Positioning an Inset Using Compass Point Values

To position an inset by using a compass point position, specify the value N, NE, E, SE, S, SW, W, or NW with 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:

data Score;
  input Student $ PreTest PostTest @@;
  label ScoreChange = 'Change in Test Scores';
  ScoreChange = PostTest - PreTest;
datalines;
Capalleti 94 91 Dubose 51 65
Engles 95 97 Grant 63 75
Krupski 80 75 Lundsford 92 55
Mcbane 75 78 Mullen 89 82
Nguyen 79 76 Patel 71 77
Si 75 70 Tanaka 87 73;

title 'Test Scores for a College Course';
ods graphics off;
proc univariate data=Score noprint;
  histogram PreTest / midpoints = 45 to 95 by 10;
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
Positioning Insets in the Margins

To position an inset in one of the four margins that surround the plot area, specify the value LM, RM, TM, or BM with the POSITION= option. 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, the inset is likely to collide with the data display.
Positioning an Inset Using Coordinates

To position an inset with coordinates, use POSITION=(x,y). You specify the coordinates in axis data units or in axis percentage units (the default). **NOTE**: You cannot position an inset with coordinates when producing ODS Graphics output.

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 45 on the horizontal axis and 10 on the vertical axis:

```plaintext
title 'Test Scores for a College Course';
ods graphics off;
proc univariate data=Score noprint;
   histogram PreTest / midpoints = 45 to 95 by 10;
   inset n / header = 'Position=(45,10)' position = (45,10) data;
run;
```

**Figure 4.12** Coordinate Position for Inset

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 below).
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:

``` SAS
title 'Test Scores for a College Course';
ods graphics off;
proc univariate data=Score noprint;
   histogram PreTest / midpoints = 45 to 95 by 10;
   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.
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 these distributions are discussed by Johnson, Kotz, and Balakrishnan (1994, 1995).

Beta Distribution

The fitted density function is

\[ p(x) = \begin{cases} 
  h \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha, \beta)\sigma^{\alpha+\beta-1}} & \text{for } \theta < x < \theta + \sigma \\
  0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma
\end{cases} \]

where \( B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \) and

A sample program for these examples, `univar3.sas`, is available in the SAS Sample Library for Base SAS software.
\( \theta = \) 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

\( v = \) vertical scaling factor

\[
v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION}
\end{cases}
\]

**NOTE:** This notation is consistent with that of other distributions that you can fit with the HISTOGRAM statement. However, many texts, including Johnson, Kotz, and Balakrishnan (1995), write the beta density function as

\[
p(x) = \begin{cases} 
  (x-a)^{p-1}(b-x)^{q-1} & \text{for } a < x < b \\
  0 & \text{for } x \leq a \text{ or } x \geq b
\end{cases}
\]

The two parameterizations are related as follows:

\[
\begin{align*}
\sigma &= b - a \\
\theta &= a \\
\alpha &= p \\
\beta &= q
\end{align*}
\]

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 by using the BETA option, \( \theta \) must be less than the minimum data value and \( \theta + \sigma \) must be greater than the maximum data value. You can specify \( \theta \) and \( \sigma \) with the THETA= and SIGMA= **beta-options** 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 \( \theta \) and \( \sigma \). However, three- and four-parameter maximum likelihood estimation does not always converge.

In addition, you can specify \( \alpha \) and \( \beta \) with the ALPHA= and BETA= **beta-options**, respectively. By default, the procedure calculates maximum likelihood estimates for \( \alpha \) and \( \beta \). 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 \( \alpha \) and \( \beta \), 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 QUANTILE to compute beta quantiles and the DATA step function CDF to compute beta probabilities.
Exponential Distribution

The fitted density function is

\[ p(x) = \begin{cases} \frac{hv}{\theta} \exp\left(-\frac{x}{\theta}\right) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases} \]

where

\[ \theta = \text{threshold parameter} \]
\[ \sigma = \text{scale parameter } (\sigma > 0) \]
\[ h = \text{width of histogram interval} \]
\[ v = \text{vertical scaling factor} \]

and

\[ v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases} \]

The threshold parameter \( \theta \) must be less than or equal to the minimum data value. You can specify \( \theta \) with the THRESHOLD= exponential-option. By default, \( \theta = 0 \). If you specify THETA=EST, a maximum likelihood estimate is computed for \( \theta \). In addition, you can specify \( \sigma \) with the SCALE= exponential-option. By default, the procedure calculates a maximum likelihood estimate for \( \sigma \). 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 \( Y = \exp(-X) \) has an exponential distribution, then \( X \) has an extreme value distribution.

You can use the DATA step function QUANTILE to compute exponential quantiles and the DATA step function CDF to compute exponential probabilities.

Gamma Distribution

The fitted density function is

\[ p(x) = \begin{cases} \frac{hv}{\Gamma(\alpha)\sigma} \left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases} \]

where

\[ \theta = \text{threshold parameter} \]
\[ \sigma = \text{scale parameter } (\sigma > 0) \]
\[ \alpha = \text{shape parameter } (\alpha > 0) \]
\[ h = \text{width of histogram interval} \]
\[ v = \text{vertical scaling factor} \]

and

\[
v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}
\]

The threshold parameter \( \theta \) must be less than the minimum data value. You can specify \( \theta \) with the THRESHOLD= gamma-option. By default, \( \theta = 0 \). If you specify THETA=EST, a maximum likelihood estimate is computed for \( \theta \). In addition, you can specify \( \sigma \) and \( \alpha \) with the SCALE= and ALPHA= gamma-options. By default, the procedure calculates maximum likelihood estimates for \( \sigma \) and \( \alpha \).

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

\[
p(x) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} \left( \frac{x}{\nu} \right)^{\frac{\nu}{2}-1} \exp \left( -\frac{x}{\nu} \right) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}
\]

Notice that this is a gamma distribution with \( \alpha = \frac{\nu}{2}, \sigma = 2, \) and \( \theta = 0 \). The exponential distribution is a gamma distribution with \( \alpha = 1 \), and the Erlang distribution is a gamma distribution with \( \alpha \) being a positive integer. A related distribution is the Rayleigh distribution. If \( R = \frac{\max(X_1, \ldots, X_\nu)}{\min(X_1, \ldots, X_\nu)} \) where the \( X_i \)'s are independent \( \chi^2 \) variables, then \( \log R \) is distributed with a \( \chi^2 \) distribution having a probability density function of

\[
p(x) = \begin{cases} \left[ 2^{\frac{\nu}{2}-1} \Gamma(\frac{\nu}{2}) \right]^{-1} \frac{x^{\nu-1} \exp(-\frac{x^2}{2})}{\nu} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}
\]

If \( \nu = 2 \), the preceding distribution is referred to as the Rayleigh distribution.

You can use the DATA step function QUANTILE to compute gamma quantiles and the DATA step function CDF to compute gamma probabilities.

**Gumbel Distribution**

The fitted density function is

\[
p(x) = \frac{h \nu}{\sigma} e^{-(x-\mu)/\sigma} \exp \left( -e^{-(x-\mu)/\sigma} \right)
\]

where

\[
\mu = \text{location parameter} \\
\sigma = \text{scale parameter (} \sigma > 0 \text{)} \\
h = \text{width of histogram interval}
\]
\( v = \text{vertical scaling factor} \)

and

\[
v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION}
\end{cases}
\]

You can specify \( \mu \) and \( \sigma \) with the MU= and SIGMA= Gumbel-options, respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.

**NOTE:** The Gumbel distribution is also referred to as Type 1 extreme value distribution.

**NOTE:** The random variable \( X \) has Gumbel (Type 1 extreme value) distribution if and only if \( e^X \) has Weibull distribution and \( \exp((X - \mu)/\sigma) \) has standard exponential distribution.

**Inverse Gaussian Distribution**

The fitted density function is

\[
p(x) = \begin{cases} 
  hv \left( \frac{\lambda}{2\pi x^3} \right)^{1/2} \exp\left( -\frac{\lambda}{2\mu^2 x} (x - \mu)^2 \right) & \text{for } x > 0 \\
  0 & \text{for } x \leq 0
\end{cases}
\]

where

\( \mu = \text{location parameter} (\mu > 0) \)

\( \lambda = \text{shape parameter} (\lambda > 0) \)

\( h = \text{width of histogram interval} \)

\( v = \text{vertical scaling factor} \)

and

\[
v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION}
\end{cases}
\]

The location parameter \( \mu \) has to be greater than zero. You can specify \( \mu \) with the MU= iGauss-option. In addition, you can specify shape parameter \( \lambda \) with LAMBDA= iGauss-option. By default, the procedure calculates maximum likelihood estimates for \( \mu \) and \( \lambda \).

**NOTE:** The special case where \( \mu = 1 \) and \( \lambda = \phi \) corresponds to the Wald distribution.

You can use the DATA step function QUANTILE to compute inverse Gaussian quantiles and the DATA step function CDF to compute inverse Gaussian probabilities.
Lognormal Distribution

The fitted density function is

\[ p(x) = \begin{cases} \frac{hv}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2} \right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases} \]

where

\[ \theta = \text{threshold parameter} \]
\[ \zeta = \text{scale parameter} \quad (-\infty < \zeta < \infty) \]
\[ \sigma = \text{shape parameter} \quad (\sigma > 0) \]
\[ h = \text{width of histogram interval} \]
\[ v = \text{vertical scaling factor} \]

and

\[ v = \begin{cases} n & \text{the sample size, for } \text{VSCALE=}\text{COUNT} \\ 100 & \text{for } \text{VSCALE=}\text{PERCENT} \\ 1 & \text{for } \text{VSCALE=}\text{PROPORTION} \end{cases} \]

The threshold parameter \( \theta \) must be less than the minimum data value. You can specify \( \theta \) with the \texttt{THRESHOLD=} \texttt{lognormal-option}. By default, \( \theta = 0 \). If you specify \texttt{THETA=EST}, a maximum likelihood estimate is computed for \( \theta \). You can specify \( \zeta \) and \( \sigma \) with the \texttt{SCALE=} and \texttt{SHAPE=} \texttt{lognormal-options}, respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.

\textbf{NOTE:} The lognormal distribution is also referred to as the \( SL \) distribution in the Johnson system of distributions.

\textbf{NOTE:} This book uses \( \sigma \) to denote the shape parameter of the lognormal distribution, whereas \( \sigma \) is used to denote the scale parameter of the other distributions. The use of \( \sigma \) 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. Based on this relationship, you can use the DATA step function \texttt{PROBIT} to compute lognormal quantiles and the DATA step function \texttt{PROBNORM} to compute probabilities.

Normal Distribution

The fitted density function is

\[ p(x) = \frac{hv}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right) \quad \text{for } -\infty < x < \infty \]

where

\[ \mu = \text{mean} \]
\[ \sigma = \text{standard deviation} \quad (\sigma > 0) \]
$$h = \text{width of histogram interval}$$

$$v = \text{vertical scaling factor}$$

and

$$v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION} 
\end{cases}$$

You can specify $\mu$ and $\sigma$ with the MU= and SIGMA= normal-options, respectively. By default, the procedure estimates $\mu$ with the sample mean and $\sigma$ with the sample standard deviation.

You can use the DATA step function QUANTILE to compute beta quantiles and the DATA step function CDF to compute normal probabilities.

**NOTE:** The normal distribution is also referred to as the $S_N$ distribution in the Johnson system of distributions.

### Generalized Pareto Distribution

The fitted density function is

$$p(x) = \begin{cases} 
  \frac{h v}{\sigma} (1 - \alpha (x - \theta) / \sigma)^{1/\alpha - 1} & \text{if } \alpha \neq 0 \\
  \frac{h v}{\sigma} \exp(-x / \sigma) & \text{if } \alpha = 0 
\end{cases}$$

where

- $\theta = \text{threshold parameter}$
- $\alpha = \text{shape parameter}$
- $\sigma = \text{shape parameter} (\sigma > 0)$
- $h = \text{width of histogram interval}$
- $v = \text{vertical scaling factor}$

and

$$v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION} 
\end{cases}$$

The support of the distribution is $x > 0$ for $\alpha \leq 0$ and $0 < x < \sigma / \alpha$ for $\alpha > 0$.

**NOTE:** Special cases of Pareto distribution with $\alpha = 0$ and $\alpha = 1$ correspond respectively to the exponential distribution with mean $\sigma$ and uniform distribution on the interval $(0, \sigma)$.

The threshold parameter $\theta$ must be less than the minimum data value. You can specify $\theta$ with the THETA= Pareto-option. By default, $\theta = 0$. You can also specify $\alpha$ and $\sigma$ with the ALPHA= and SIGMA= Pareto-options, respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.
**NOTE:** Maximum likelihood estimation of the parameters works well if $\alpha \leq \frac{1}{2}$, but not otherwise. In this case the estimators are asymptotically normal and asymptotically efficient. The asymptotic normal distribution of the maximum likelihood estimates has mean $(\alpha, \sigma)$ and variance-covariance matrix

$$
\frac{1}{n} \begin{pmatrix}
(1 - \alpha)^2 & \sigma (1 - \alpha) \\
\sigma (1 - \alpha) & 2 \sigma^2 (1 - \alpha)
\end{pmatrix}.
$$

**NOTE:** If no local minimum found in the space

$$\{\alpha < 0, \sigma > 0\} \cup \{0 < \alpha \leq 1, \sigma / \alpha > \max(X_i)\},$$

there is no maximum likelihood estimator. More details on how to find maximum likelihood estimators and suggested algorithm can be found in Grimshaw(1993).

**Power Function Distribution**

The fitted density function is

$$p(x) = \begin{cases} 
  hv \frac{\alpha}{\sigma} \left(\frac{x - \theta}{\sigma}\right)^{\alpha-1} & \text{for } \theta < x < \theta + \sigma \\
  0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma
\end{cases}$$

where

- $\theta = \text{lower threshold parameter (lower endpoint parameter)}$
- $\sigma = \text{scale parameter (} \sigma > 0\text{)}$
- $\alpha = \text{shape parameter (} \alpha > 0\text{)}$
- $h = \text{width of histogram interval}$
- $v = \text{vertical scaling factor}$

and

$$v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION}
\end{cases}$$

**NOTE:** This notation is consistent with that of other distributions that you can fit with the HISTOGRAM statement. However, many texts, including Johnson, Kotz, and Balakrishnan (1995), write the density function of power function distribution as

$$p(x) = \begin{cases} 
  \frac{p}{b-a} \left(\frac{x-a}{b-a}\right)^{p-1} & \text{for } a < x < b \\
  0 & \text{for } x \leq a \text{ or } x \geq b
\end{cases}$$

The two parameterizations are related as follows:

$$\sigma = b - a$$
\[ \theta = a \]
\[ \alpha = p \]

**NOTE:** The family of power function distributions is subclass of beta distribution with density function

\[ p(x) = \begin{cases} 
  hv \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha, \beta)} & \text{for } \theta < x < \theta + \sigma \\
  0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma 
\end{cases} \]

where \( B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \) with parameter \( \beta = 1 \). Therefore, all properties and estimation procedures of beta distribution apply.

The range of the power function distribution is bounded below by a threshold parameter \( \theta = a \) and above by \( \theta + \sigma = b \). If you specify a fitted power function curve by using the POWER option, \( \theta \) must be less than the minimum data value and \( \theta + \sigma \) must be greater than the maximum data value. You can specify \( \theta \) and \( \sigma \) with the THETA= and SIGMA= power-options in parentheses after the keyword POWER. By default, \( \sigma = 1 \) and \( \theta = 0 \). If you specify THETA=EST and SIGMA=EST, maximum likelihood estimates are computed for \( \theta \) and \( \sigma \). However, three-parameter maximum likelihood estimation does not always converge.

In addition, you can specify \( \alpha \) with the ALPHA= power-option. By default, the procedure calculates maximum likelihood estimate for \( \alpha \). For example, to fit a power function density curve to a set of data bounded below by 32 and above by 212 with maximum likelihood estimate for \( \alpha \), use the following statement:

```
histogram Length / power(theta=32 sigma=180);
```

**Rayleigh Distribution**

The fitted density function is

\[ p(x) = \begin{cases} 
  hv \frac{x-\theta}{\sigma^2} e^{-(x-\theta)^2/(2\sigma^2)} & \text{for } x \geq \theta \\
  0 & \text{for } x < \theta 
\end{cases} \]

where

- \( \theta = \) lower threshold parameter (lower endpoint parameter)
- \( \sigma = \) scale parameter (\( \sigma > 0 \))
- \( h = \) width of histogram interval
- \( v = \) vertical scaling factor

and

\[ v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION} 
\end{cases} \]

**NOTE:** The Rayleigh distribution is Weibull distribution with density function

\[ p(x) = \begin{cases} 
  hv \frac{k}{\lambda} \left( \frac{x-\theta}{\lambda} \right)^{k-1} \exp\left(-\left( \frac{x-\theta}{\lambda} \right)^k\right) & \text{for } x \geq \theta \\
  0 & \text{for } x < \theta 
\end{cases} \]

and with shape parameter \( k = 2 \) and scale parameter \( \lambda = \sqrt{2} \sigma \).
The threshold parameter $\theta$ must be less than the minimum data value. You can specify $\theta$ with the THETA=Rayleigh-option. By default, $\theta = 0$. In addition you can specify $\sigma$ with the SIGMA=Rayleigh-option. By default, the procedure calculates maximum likelihood estimate for $\sigma$.

For example, to fit a Rayleigh density curve to a set of data bounded below by 32 with maximum likelihood estimate for $\sigma$, use the following statement:

```
histogram Length / rayleigh(theta=32);
```

**Johnson $S_B$ Distribution**

The fitted density function is

$$p(x) = \begin{cases} 
\frac{\delta h v}{\sigma^2 \pi} \left[ \left( \frac{x-\theta}{\sigma} \right) \left( 1 - \frac{x-\theta}{\sigma} \right) \right]^{-1} \times \\
\exp \left[ -\frac{1}{2} \left( \gamma + \delta \log \left( \frac{x-\theta}{\sigma + x} \right) \right)^2 \right] & \text{for } \theta < x < \theta + \sigma \\
0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma
\end{cases}$$

for $\theta < x < \theta + \sigma$

where

- $\theta = \text{threshold parameter} \ (\ -\infty < \theta < \infty )$
- $\sigma = \text{scale parameter} \ (\sigma > 0)$
- $\delta = \text{shape parameter} \ (\delta > 0)$
- $\gamma = \text{shape parameter} \ (-\infty < \gamma < \infty )$
- $h = \text{width of histogram interval}$
- $v = \text{vertical scaling factor}$

and

$$v = \begin{cases} 
n & \text{the sample size, for VSCALE=COUNT} \\
100 & \text{for VSCALE=PERCENT} \\
1 & \text{for VSCALE=PROPORTION}
\end{cases}$$

The $S_B$ distribution is bounded below by the parameter $\theta$ and above by the value $\theta + \sigma$. The parameter $\theta$ must be less than the minimum data value. You can specify $\theta$ with the THETA= $S_B$-option, or you can request that $\theta$ be estimated with the THETA = EST $S_B$-option. The default value for $\theta$ is zero. The sum $\theta + \sigma$ must be greater than the maximum data value. The default value for $\sigma$ is one. You can specify $\sigma$ with the SIGMA= $S_B$-option, or you can request that $\sigma$ be estimated with the SIGMA = EST $S_B$-option.

By default, the method of percentiles given by Slifker and Shapiro (1980) is used to estimate the parameters. This method is based on four data percentiles, denoted by $x_{-3z}$, $x_{-z}$, $x_z$, and $x_{3z}$, which correspond to the four equally spaced percentiles of a standard normal distribution, denoted by $-3z$, $-z$, $z$, and $3z$, under the transformation

$$z = \gamma + \delta \log \left( \frac{x-\theta}{\theta + \sigma - x} \right)$$
The default value of $z$ is 0.524. The results of the fit are dependent on the choice of $z$, and you can specify other values with the FITINTERVAL= option (specified in parentheses after the SB option). If you use the method of percentiles, you should select a value of $z$ that corresponds to percentiles which are critical to your application.

The following values are computed from the data percentiles:

\[
\begin{align*}
    m &= x_{3z} - x_z \\
n &= x_{-z} - x_{-3z} \\
p &= x_z - x_{-z}
\end{align*}
\]

It was demonstrated by Slifker and Shapiro (1980) that

\[
\begin{align*}
    \frac{mn}{p^2} > 1 & \quad \text{for any } S_U \text{ distribution} \\
    \frac{mn}{p^2} < 1 & \quad \text{for any } S_B \text{ distribution} \\
    \frac{mn}{p^2} = 1 & \quad \text{for any } S_L \text{ (lognormal) distribution}
\end{align*}
\]

A tolerance interval around one is used to discriminate among the three families with this ratio criterion. You can specify the tolerance with the FITTOLERANCE= option (specified in parentheses after the SB option). The default tolerance is 0.01. Assuming that the criterion satisfies the inequality

\[
\frac{mn}{p^2} < 1 - \text{tolerance}
\]

the parameters of the $S_B$ distribution are computed using the explicit formulas derived by Slifker and Shapiro (1980).

If you specify FITMETHOD = MOMENTS (in parentheses after the SB option), the method of moments is used to estimate the parameters. If you specify FITMETHOD = MLE (in parentheses after the SB option), the method of maximum likelihood is used to estimate the parameters. Note that maximum likelihood estimates may not always exist. Refer to Bowman and Shenton (1983) for discussion of methods for fitting Johnson distributions.

**Johnson $S_U$ Distribution**

The fitted density function is

\[
p(x) = \begin{cases} 
    \frac{\delta hv}{\sigma^2 \sqrt{2\pi}} \frac{1}{\sqrt{1 + (x-\theta)/\sigma}^2} x \\
    \exp \left[ -\frac{1}{2} \left( \gamma + \delta \sinh^{-1} \left( \frac{x-\theta}{\sigma} \right) \right)^2 \right] & \text{for } x > \theta \\
    0 & \text{for } x \leq \theta
\end{cases}
\]

where

\[
\begin{align*}
    \theta &= \text{location parameter } (-\infty < \theta < \infty) \\
    \sigma &= \text{scale parameter } (\sigma > 0) \\
    \delta &= \text{shape parameter } (\delta > 0) \\
    \gamma &= \text{shape parameter } (-\infty < \gamma < \infty)
\end{align*}
\]
\[ h = \text{width of histogram interval} \]
\[ v = \text{vertical scaling factor} \]

and

\[ v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION} 
\end{cases} \]

You can specify the parameters with the THETA=, SIGMA=, DELTA=, and GAMMA= options, which are enclosed in parentheses after the SU option. If you do not specify these parameters, they are estimated.

By default, the method of percentiles given by Slifker and Shapiro (1980) is used to estimate the parameters. This method is based on four data percentiles, denoted by \( x_{3z}, x_{-z}, x_z, \) and \( x_{3z}, \) which correspond to the four equally spaced percentiles of a standard normal distribution, denoted by \(-3z, -z, z, \) and \( 3z, \) under the transformation

\[ z = \gamma + \delta \sinh^{-1} \left( \frac{x - \theta}{\sigma} \right) \]

The default value of \( z \) is 0.524. The results of the fit are dependent on the choice of \( z, \) and you can specify other values with the FITINTERVAL= option (specified in parentheses after the SU option). If you use the method of percentiles, you should select a value of \( z \) that corresponds to percentiles that are critical to your application.

The following values are computed from the data percentiles:

\[
\begin{align*}
  m &= x_{3z} - x_z \\
  n &= x_{-z} - x_{-3z} \\
  p &= x_z - x_{-z}
\end{align*}
\]

It was demonstrated by Slifker and Shapiro (1980) that

\[
\frac{mn}{p^2} > 1 \quad \text{for any} \ S_U \ \text{distribution} \\
\frac{mn}{p^2} < 1 \quad \text{for any} \ S_B \ \text{distribution} \\
\frac{mn}{p^2} = 1 \quad \text{for any} \ S_L \ (\text{lognormal}) \ \text{distribution}
\]

A tolerance interval around one is used to discriminate among the three families with this ratio criterion. You can specify the tolerance with the FITTOLERANCE= option (specified in parentheses after the SU option). The default tolerance is 0.01. Assuming that the criterion satisfies the inequality

\[ \frac{mn}{p^2} > 1 + \text{tolerance} \]

the parameters of the \( S_U \) distribution are computed using the explicit formulas derived by Slifker and Shapiro (1980).

If you specify FITMETHOD = MOMENTS (in parentheses after the SU option), the method of moments is used to estimate the parameters. If you specify FITMETHOD = MLE (in parentheses after the SU option), the method of maximum likelihood is used to estimate the parameters. Note that maximum likelihood estimates do not always exist. Refer to Bowman and Shenton (1983) for discussion of methods for fitting Johnson distributions.
**Weibull Distribution**

The fitted density function is

\[
p(x) = \begin{cases} 
  hv \frac{x-\theta}{\sigma}^{c-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right) & \text{for } x > \theta \\
  0 & \text{for } x \leq \theta 
\end{cases}
\]

where

- \(\theta\) = threshold parameter
- \(\sigma\) = scale parameter (\(\sigma > 0\))
- \(c\) = shape parameter (\(c > 0\))
- \(h\) = width of histogram interval
- \(v\) = vertical scaling factor

and

\[
v = \begin{cases} 
  n & \text{the sample size, for VSCALE=COUNT} \\
  100 & \text{for VSCALE=PERCENT} \\
  1 & \text{for VSCALE=PROPORTION}
\end{cases}
\]

The threshold parameter \(\theta\) must be less than the minimum data value. You can specify \(\theta\) with the THRESHOLD= **Weibull-option**. By default, \(\theta = 0\). If you specify THETA=EST, a maximum likelihood estimate is computed for \(\theta\). You can specify \(\sigma\) and \(c\) with the SCALE= and SHAPE= **Weibull-options**, respectively. By default, the procedure calculates maximum likelihood estimates for \(\sigma\) and \(c\).

The exponential distribution is a special case of the Weibull distribution where \(c = 1\).

You can use the DATA step function QUANTILE to compute Weibull quantiles and the DATA step function CDF to compute Weibull probabilities.

---

**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 goodness-of-fit tests for the null hypothesis that the values of the analysis variable are a random sample from the specified theoretical distribution. See Example 4.22.

When you specify the NORMAL option, these tests, which are summarized in the output table labeled “Tests for Normality,” include the following:

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Anderson-Darling test
- Cramér–von Mises test
The Kolmogorov-Smirnov \( D \) statistic, the Anderson-Darling statistic, and the Cramér–von Mises statistic are based on the empirical distribution function (EDF). However, some EDF tests are not supported when certain combinations of the parameters of a specified distribution are estimated. See Table 4.32 for a list of the EDF tests available. You determine whether to reject the null hypothesis by examining the \( p \)-value that is associated with a goodness-of-fit statistic. When the \( p \)-value is less than the predetermined critical value (\( \alpha \)), you reject the null hypothesis and conclude that the data did not come from the specified distribution.

If you want to test the normality assumptions for 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. Because 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, the PROBPLOT statement, and the 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 can 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, \( W \) (also denoted as \( W_n \) to emphasize its dependence on the sample size \( n \)). 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 and Wilk 1965). When \( n \) is greater than three, the coefficients to compute the linear combination of the order statistics are approximated by the method of Royston (1992). The statistic \( W \) is always greater than zero and less than or equal to one \( (0 < W \leq 1) \).

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. The method for computing the \( p \)-value (the probability of obtaining a \( W \) statistic less than or equal to the observed value) depends on \( n \). For \( n = 3 \), the probability distribution of \( W \) is known and is used to determine the \( p \)-value. For \( n > 4 \), a normalizing transformation is computed:

\[
Z_n = \begin{cases} 
(-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \leq n \leq 11 \\
(\log(1 - W_n) - \mu)/\sigma & \text{if } 12 \leq n \leq 2000 
\end{cases}
\]

The values of \( \sigma, \gamma, \) and \( \mu \) are functions of \( n \) obtained from simulation results. Large values of \( Z_n \) indicate departure from normality, and because the statistic \( Z_n \) has an approximately standard normal distribution, this distribution is used to determine the \( p \)-values for \( n > 4 \).

**EDF Goodness-of-Fit Tests**

When you fit a parametric distribution, PROC UNIVARIATE provides a series of goodness-of-fit tests based on the empirical distribution function (EDF). The EDF tests offer advantages over traditional chi-square
goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints. For a thorough discussion, refer to D’Agostino and Stephens (1986).

The empirical distribution function is defined for a set of \( n \) independent observations \( X_1, \ldots, X_n \) with a common distribution function \( F(x) \). Denote the observations ordered from smallest to largest as \( X_{(1)}, \ldots, X_{(n)} \). The empirical distribution function, \( F_n(x) \), is defined as

\[
F_n(x) = 0, \quad x < X_{(1)} \\
F_n(x) = \frac{i}{n}, \quad X_{(i)} \leq x < X_{(i+1)} \quad i = 1, \ldots, n - 1 \\
F_n(x) = 1, \quad X_{(n)} \leq x
\]

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 less than or equal to \( x \), while \( F(x) \) is the probability of an observation 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 make use of 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 \( X_{(1)}, \ldots, X_{(n)} \), the values \( U_{(i)} = F(X_{(i)}) \) are computed by applying the transformation, as discussed in the next three sections.

PROC UNIVARIATE provides three EDF tests:

- Kolmogorov-Smirnov
- Anderson-Darling
- Cramér–von Mises

The following sections provide formal definitions of these EDF statistics.

**Kolmogorov D Statistic**

The Kolmogorov-Smirnov statistic (\( D \)) is defined as

\[
D = \sup_x |F_n(x) - F(x)|
\]

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 \( D^- \), where \( D^+ \) is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function, and \( D^- \) is the largest vertical distance when the EDF is less than the distribution function.

\[
D^+ = \max_i \left( \frac{i}{n} - U_{(i)} \right) \\
D^- = \max_i \left( U_{(i)} - \frac{i-1}{n} \right) \\
D = \max(D^+, D^-)
\]

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 Cramér–von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference \((F_n(x) - F(x))^2\). Quadratic statistics have the following general form:

\[
Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \psi(x) dF(x)
\]

The function \(\psi(x)\) weights the squared difference \((F_n(x) - F(x))^2\).

The Anderson-Darling statistic \((A^2)\) is defined as

\[
A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 [F(x) (1 - F(x))]^{-1} dF(x)
\]

Here the weight function is \(\psi(x) = [F(x) (1 - F(x))]^{-1}\).

The Anderson-Darling statistic is computed as

\[
A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[ (2i - 1) \log U(i) + (2n + 1 - 2i) \log(1 - U(i)) \right]
\]

**Cramér–von Mises Statistic**

The Cramér–von Mises statistic \((W^2)\) is defined as

\[
W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 dF(x)
\]

Here the weight function is \(\psi(x) = 1\).

The Cramér–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 \((p\)-values\). For the Gumbel, inverse Gaussian, generalized Pareto, and Rayleigh distributions, PROC UNIVARIATE computes associated probability values \((p\)-values\) by resampling from the estimated distribution. By default 500 EDF test statistics are computed and then compared to the EDF test statistic for the specified (fitted) distribution. The number of samples can be controlled by setting EDFNSAMPLES=\(n\). For example, to request Gumbel distribution Goodness-of-Fit test \(p\)-values based on 5000 simulations, use the following statement:
proc univariate data=test;
  histogram / gumbel(edfnsamples=5000);
run;

For the beta, exponential, gamma, lognormal, normal, power function, and Weibull distributions the UNI-
VARIATE procedure uses internal tables of probability levels similar to those given by D’Agostino and
Stephens (1986). If the value is between two probability levels, then linear interpolation is used to estimate
the probability value.

The probability value depends upon the parameters that are known and the parameters that are estimated for
the distribution. Table 4.32 summarizes different combinations fitted for which EDF tests are available.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Threshold</th>
<th>Parameters</th>
<th>Tests Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta</td>
<td>θ known</td>
<td>σ known</td>
<td>α, β known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ known</td>
<td>α, β &lt; 5 unknown</td>
</tr>
<tr>
<td>exponential</td>
<td>θ known,</td>
<td>σ known</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ known</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ unknown</td>
<td></td>
</tr>
<tr>
<td>gamma</td>
<td>θ known</td>
<td>σ known</td>
<td>α known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ unknown</td>
<td>α known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ known</td>
<td>α unknown</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ unknown</td>
<td>α &gt; 1 unknown</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ known</td>
<td>α &gt; 1 known</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ unknown</td>
<td>α &gt; 1 known</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ known</td>
<td>α &gt; 1 unknown</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ unknown</td>
<td>α &gt; 1 unknown</td>
</tr>
<tr>
<td>lognormal</td>
<td>θ known</td>
<td>ζ known</td>
<td>σ known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>ζ known</td>
<td>σ unknown</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>ζ unknown</td>
<td>σ known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>ζ unknown</td>
<td>σ unknown</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>ζ known</td>
<td>σ &lt; 3 known</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>ζ known</td>
<td>σ &lt; 3 unknown</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>ζ unknown</td>
<td>σ &lt; 3 known</td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>ζ unknown</td>
<td>σ &lt; 3 unknown</td>
</tr>
<tr>
<td>normal</td>
<td>θ known</td>
<td>σ known</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ known</td>
<td></td>
</tr>
<tr>
<td></td>
<td>θ unknown</td>
<td>σ unknown</td>
<td></td>
</tr>
<tr>
<td>power function</td>
<td>θ known</td>
<td>σ known</td>
<td>α known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ known</td>
<td>α &lt; 5 unknown</td>
</tr>
<tr>
<td>Weibull</td>
<td>θ known</td>
<td>σ known</td>
<td>c known</td>
</tr>
<tr>
<td></td>
<td>θ known</td>
<td>σ unknown</td>
<td>c known</td>
</tr>
</tbody>
</table>

Table 4.32 Availabilty of EDF Tests
### Table 4.32 (continued)

<table>
<thead>
<tr>
<th>Distribution Parameters</th>
<th>Threshold</th>
<th>Tests Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold Scale Shape</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta$ known $\sigma$ known $c$ unknown</td>
<td>$A^2$ and $W^2$</td>
<td></td>
</tr>
<tr>
<td>$\theta$ known $\sigma$ unknown $c$ unknown</td>
<td>$A^2$ and $W^2$</td>
<td></td>
</tr>
<tr>
<td>$\theta$ unknown $\sigma$ known $c &gt; 2$ known</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>$\theta$ unknown $\sigma$ unknown $c &gt; 2$ known</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>$\theta$ unknown $\sigma$ known $c &gt; 2$ unknown</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>$\theta$ unknown $\sigma$ unknown $c &gt; 2$ unknown</td>
<td>all</td>
<td></td>
</tr>
</tbody>
</table>

---

**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 identify features that might be obscured by the choice of histogram bins or sampling variation. A kernel density estimate can also be more effective than a parametric curve fit when the process distribution is multi-modal. See Example 4.23.

The general form of the kernel density estimator is

$$\hat{f}_h(x) = \frac{h}{n\lambda} \sum_{i=1}^{n} K_0 \left( \frac{x - x_i}{\lambda} \right)$$

where

- $K_0(\cdot)$ is the kernel function
- $\lambda$ is the bandwidth
- $n$ is the sample size
- $x_i$ is the $i$th observation
- $v = \text{vertical scaling factor}$

and

$$v = \begin{cases} 
100 & \text{for VSCALE=PERCENT} \\
1 & \text{for VSCALE=PROPORTION} \\
n & \text{for VSCALE=COUNT}
\end{cases}$$

The KERNEL option provides three kernel functions ($K_0$): normal, quadratic, and triangular. You can specify the function with the K= kernel-option 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: $K_0(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)$ for $-\infty < t < \infty$
- Quadratic: $K_0(t) = \frac{3}{4}(1 - t^2)$ for $|t| \leq 1$
- Triangular: $K_0(t) = 1 - |t|$ for $|t| \leq 1$
The value of \( \lambda \), referred to as the bandwidth parameter, determines the degree of smoothness in the estimated density function. You specify \( \lambda \) indirectly by specifying a standardized bandwidth \( c \) with the \texttt{C=} kernel option. If \( Q \) is the interquartile range and \( n \) is the sample size, then \( c \) is related to \( \lambda \) by the formula

\[
\lambda = c Q n^{-\frac{1}{5}}
\]

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):

\[
\text{MISE}(\lambda) = \int_x \left\{ E(\hat{f}_\lambda(x)) - f(x) \right\}^2 dx + \int_x \text{var}(\hat{f}_\lambda(x)) dx
\]

The MISE is the sum of the integrated squared bias and the variance. An approximate mean integrated square error (AMISE) is:

\[
\text{AMISE}(\lambda) = \frac{1}{4} \lambda^4 \left( \int_t t^2 K(t) dt \right)^2 \int_x \left( f''(x) \right)^2 dx + \frac{1}{n\lambda} \int_t K(t)^2 dt
\]

A bandwidth that minimizes AMISE can be derived by treating \( f(x) \) as the normal density that has parameters \( \mu \) and \( \sigma \) estimated by the sample mean and standard deviation. If you do not specify a bandwidth parameter or if you specify \texttt{C=MISE}, the bandwidth that minimizes AMISE is used. The value of AMISE can be used to compare different density estimates. You can also specify \texttt{C=SJPI} to select the bandwidth by using a plug-in formula of Sheather and Jones (Jones, Marron, and Sheather 1996). For each estimate, the bandwidth parameter \( c \), the kernel function type, and the value of AMISE are reported in the SAS log.

The general kernel density estimates assume that the domain of the density to estimate can take on all values on a real line. However, sometimes the domain of a density is an interval bounded on one or both sides. For example, if a variable \( Y \) is a measurement of only positive values, then the kernel density curve should be bounded so that is zero for negative \( Y \) values. You can use the \texttt{LOWER=} and \texttt{UPPER=} kernel options to specify the bounds.

The \texttt{UNIVARIATE} procedure uses a reflection technique to create the bounded kernel density curve, as described in Silverman (1986, pp. 30-31). It adds the reflections of the kernel density that are outside the boundary to the bounded kernel estimates. The general form of the bounded kernel density estimator is computed by replacing \( K_0 \left( \frac{x-x_i}{\lambda} \right) \) in the original equation with

\[
\left\{ K_0 \left( \frac{x-x_i}{\lambda} \right) + K_0 \left( \frac{(x-x_i)+(x_i-x_l)}{\lambda} \right) + K_0 \left( \frac{(x-u-x)+(x_u-x_i)}{\lambda} \right) \right\}
\]

where \( x_l \) is the lower bound and \( x_u \) is the upper bound.

Without a lower bound, \( x_l = -\infty \) and \( K_0 \left( \frac{(x-x_i)+(x_i-x_l)}{\lambda} \right) = 0 \). Similarly, without an upper bound, \( x_u = \infty \) and \( K_0 \left( \frac{(x_u-x)+(x_u-x_i)}{\lambda} \right) = 0 \).

When \texttt{C=MISE} is used with a bounded kernel density, the \texttt{UNIVARIATE} procedure uses a bandwidth that minimizes the AMISE for its corresponding unbounded kernel.
Construction of Quantile-Quantile and Probability Plots

Figure 4.14 illustrates how a Q-Q plot is constructed for a specified theoretical distribution. First, the \( n \) nonmissing values of the variable are ordered from smallest to largest:

\[
x(1) \leq x(2) \leq \cdots \leq x(n)
\]

Then the \( i \)th ordered value \( x(i) \) is plotted as 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 \( F(\cdot) \) is the specified distribution with zero location parameter and unit scale parameter.

You can modify the adjustment constants --0.375 and 0.25 with the RANKADJ= and NADJ= options. This default combination is recommended by Blom (1958). For additional information, see Chambers et al. (1983). Because \( x(i) \) 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 (see the section “PROBPLOT Statement” on page 345) are constructed the same way, except that the \( x \)-axis is scaled nonlinearly in percentiles.

**Figure 4.14** Construction of a Q-Q Plot

Interpretation of Quantile-Quantile and Probability Plots

The following properties of Q-Q plots and probability plots make them useful diagnostics of 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 Table 4.33.

In some applications, a nonlinear pattern may be more revealing than a linear pattern. However, Chambers et al. (1983) note that departures from linearity can also be due to chance variation.

<table>
<thead>
<tr>
<th>Description of Point Pattern</th>
<th>Possible Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>all but a few points fall on a line</td>
<td>outliers in the data</td>
</tr>
<tr>
<td>left end of pattern is below the line; right end of pattern is above the line</td>
<td>long tails at both ends of the data distribution</td>
</tr>
<tr>
<td>left end of pattern is above the line; right end of pattern is below the line</td>
<td>short tails at both ends of the data distribution</td>
</tr>
<tr>
<td>curved pattern with slope increasing from left to right</td>
<td>data distribution is skewed to the right</td>
</tr>
<tr>
<td>curved pattern with slope decreasing from left to right</td>
<td>data distribution is skewed to the left</td>
</tr>
<tr>
<td>staircase pattern (plateaus and gaps)</td>
<td>data have been rounded or are discrete</td>
</tr>
</tbody>
</table>

When the pattern is linear, you can use Q-Q plots to estimate shape, location, and scale parameters and to estimate percentiles. See Example 4.26 through Example 4.34.

### Distributions for Probability and Q-Q Plots

You can use the PROBPLOT and QQPLOT statements to request probability and Q-Q plots that are based on the theoretical distributions summarized in Table 4.34.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Density Function ( p(x) )</th>
<th>Range</th>
<th>Location</th>
<th>Scale</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta</td>
<td>( \frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{B(\alpha, \beta)\sigma^{\alpha+\beta-1}} )</td>
<td>( \theta &lt; x &lt; \theta + \sigma )</td>
<td>( \theta )</td>
<td>( \sigma )</td>
<td>( \alpha, \beta )</td>
</tr>
</tbody>
</table>
You can request these distributions with the BETA, EXPONENTIAL, GAMMA, PARETO, GUMBEL, LOGNORMAL, NORMAL, POWER, RAYLEIGH, WEIBULL, and WEIBULL2 options, respectively. If you do not specify a distribution option, a normal probability plot or a normal Q-Q plot is created.

The following sections provide 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 the plot, the observations are ordered from smallest to largest, and the $i$th ordered observation is plotted against the quantile $B_{\alpha\beta}^{-1} \left( \frac{i-0.375}{n+0.25} \right)$, where $B_{\alpha\beta}^{-1} (\cdot)$ is the inverse normalized incomplete beta function, $n$ is the number of nonmissing observations, and $\alpha$ and $\beta$ are the shape parameters of the beta distribution. In a probability plot, the horizontal axis is scaled in percentile units.

The pattern on the plot for $\text{ALPHA}=\alpha$ and $\text{BETA}=\beta$ tends to be linear with intercept $\theta$ and slope $\sigma$ if the data are beta distributed with the specific density function

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\theta+x-\beta)^{\beta-1}}{B(\alpha,\beta)\sigma^{\alpha+\beta-1}} & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma \end{cases}$$
where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

\[
\theta = \text{lower threshold parameter} \\
\sigma = \text{scale parameter} \ (\sigma > 0) \\
\alpha = \text{first shape parameter} \ (\alpha > 0) \\
\beta = \text{second shape parameter} \ (\beta > 0)
\]

**Exponential Distribution**

To create the plot, the observations are ordered from smallest to largest, and the $i$th ordered observation is plotted against the quantile $-\log\left(1 - \frac{i-0.375}{n+0.25}\right)$, where $n$ is the number of nonmissing observations. In a probability plot, the horizontal axis is scaled in percentile units.

The pattern on the plot tends to be linear with intercept $\theta$ and slope $\sigma$ if the data are exponentially distributed with the specific density function

\[
p(x) = \begin{cases} \\
\frac{1}{\sigma} \exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x \geq \theta \\
0 & \text{for } x < \theta
\end{cases}
\]

where $\theta$ is a threshold parameter, and $\sigma$ is a positive scale parameter.

**Gamma Distribution**

To create the plot, the observations are ordered from smallest to largest, and the $i$th ordered observation is plotted against the quantile $G_{\alpha}^{-1} \left(\frac{i-0.375}{n+0.25}\right)$, where $G_{\alpha}^{-1} (\cdot)$ is the inverse normalized incomplete gamma function, $n$ is the number of nonmissing observations, and $\alpha$ is the shape parameter of the gamma distribution. In a probability plot, the horizontal axis is scaled in percentile units.

The pattern on the plot for $\text{ALPHA} = \alpha$ tends to be linear with intercept $\theta$ and slope $\sigma$ if the data are gamma distributed with the specific density function

\[
p(x) = \begin{cases} \\
\frac{1}{\sigma \Gamma(\alpha)} \left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x > \theta \\
0 & \text{for } x \leq \theta
\end{cases}
\]

where

\[
\theta = \text{threshold parameter} \\
\sigma = \text{scale parameter} \ (\sigma > 0) \\
\alpha = \text{shape parameter} \ (\alpha > 0)
\]
Gumbel Distribution

To create the plot, the observations are ordered from smallest to largest, and the \(i\)th ordered observation is plotted against the quantile \(-\log\left(-\log\left(\frac{i-0.375}{n+0.25}\right)\right)\), where \(n\) is the number of nonmissing observations. In a probability plot, the horizontal axis is scaled in percentile units.

The pattern on the plot tends to be linear with intercept \(\mu\) and slope \(\sigma\) if the data are Gumbel distributed with the specific density function

\[
p(x) = \frac{e^{-(x-\mu)/\sigma}}{\sigma} \exp\left(-e^{-(x-\mu)/\sigma}\right)
\]

\(\mu = \) location parameter

\(\sigma = \) scale parameter (\(\sigma > 0\))

Lognormal Distribution

To create the plot, the observations are ordered from smallest to largest, and the \(i\)th ordered observation is plotted against the quantile \(\exp\left(\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)\right)\), where \(\Phi^{-1}\) is the inverse cumulative standard normal distribution, \(n\) is the number of nonmissing observations, and \(\sigma\) is the shape parameter of the lognormal distribution. In a probability plot, the horizontal axis is scaled in percentile units.

The pattern on the plot for \(\text{SIGMA}=\sigma\) tends to be linear with intercept \(\theta\) 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

\(\theta = \) threshold parameter

\(\zeta = \) scale parameter

\(\sigma = \) shape parameter (\(\sigma > 0\))

See Example 4.26 and Example 4.33.

Normal Distribution

To create the plot, the observations are ordered from smallest to largest, and the \(i\)th ordered observation is plotted 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. In a probability plot, the horizontal axis is scaled in percentile units.

The point pattern on the plot tends to be linear with intercept \(\mu\) and slope \(\sigma\) 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 \(\mu\) is the mean and \(\sigma\) is the standard deviation (\(\sigma > 0\)).
**Generalized Pareto Distribution**

To create the plot, the observations are ordered from smallest to largest, and the \( i \)th ordered observation is plotted against the quantile \((1 - (1 - \frac{i - 0.375}{n + 0.25})^\alpha) / \alpha \) (\( \alpha \neq 0 \)) or \(-\log(1 - \frac{i - 0.375}{n + 0.25}) \) (\( \alpha = 0 \)), where \( n \) is the number of nonmissing observations and \( \alpha \) is the shape parameter of the generalized Pareto distribution. The horizontal axis is scaled in percentile units.

The point pattern on the plot for \( \text{ALPHA} = \alpha \) tends to be linear with intercept \( \theta \) and slope \( \sigma \) if the data are generalized Pareto distributed with the specific density function

\[
p(x) = \begin{cases} 
\frac{1}{\alpha} (1 - \alpha(x - \theta)/\alpha)^{1/\alpha - 1} & \text{if } \alpha \neq 0 \\
\frac{1}{\sigma} \exp((-x - \theta)/\sigma) & \text{if } \alpha = 0 
\end{cases}
\]

where

- \( \theta \) = threshold parameter
- \( \sigma \) = scale parameter (\( \sigma > 0 \))
- \( \alpha \) = shape parameter (\( \alpha > 0 \))

**Power Function Distribution**

To create the plot, the observations are ordered from smallest to largest, and the \( i \)th ordered observation is plotted against the quantile \( B_{\alpha(1)}^{-1} \left( \frac{i - 0.375}{n + 0.25} \right) \), where \( B_{\alpha(1)}^{-1}() \) is the inverse normalized incomplete beta function, \( n \) is the number of nonmissing observations, \( \alpha \) is one shape parameter of the beta distribution, and the second shape parameter, \( \beta = 1 \). The horizontal axis is scaled in percentile units.

The point pattern on the plot for \( \text{ALPHA} = \alpha \) tends to be linear with intercept \( \theta \) and slope \( \sigma \) if the data are power function distributed with the specific density function

\[
p(x) = \begin{cases} 
\frac{\alpha}{\sigma} \left( \frac{x - \theta}{\sigma} \right)^{\alpha - 1} & \text{for } \theta < x < \theta + \sigma \\
0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma 
\end{cases}
\]

where

- \( \theta \) = threshold parameter
- \( \sigma \) = shape parameter (\( \sigma > 0 \))
- \( \alpha \) = shape parameter (\( \alpha > 0 \))

**Rayleigh Distribution**

To create the plot, the observations are ordered from smallest to largest, and the \( i \)th ordered observation is plotted against the quantile \( \sqrt{-2 \log \left( 1 - \frac{i - 0.375}{n + 0.25} \right)} \), where \( n \) is the number of nonmissing observations. The horizontal axis is scaled in percentile units.

The point pattern on the plot tends to be linear with intercept \( \theta \) and slope \( \sigma \) if the data are Rayleigh distributed with the specific density function

\[
p(x) = \begin{cases} 
\frac{x - \theta}{\sigma^2} \exp\left(-\frac{(x - \theta)^2}{2\sigma^2}\right) & \text{for } x \geq \theta \\
0 & \text{for } x < \theta 
\end{cases}
\]

where \( \theta \) is a threshold parameter, and \( \sigma \) is a positive scale parameter.
Three-Parameter Weibull Distribution

To create the plot, the observations are ordered from smallest to largest, and the \( i \)th ordered observation is plotted against the quantile \((-\log\left(1 - \frac{i-0.375}{n+0.25}\right))^{\frac{1}{c}}\), where \( n \) is the number of nonmissing observations, and \( c \) is the Weibull distribution shape parameter. In a probability plot, the horizontal axis is scaled in percentile units.

The pattern on the plot for \( \text{C}=c \) tends to be linear with intercept \( \theta \) and slope \( \sigma \) 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

\[
\begin{align*}
\theta &= \text{threshold parameter} \\
\sigma &= \text{scale parameter } (\sigma > 0) \\
c &= \text{shape parameter } (c > 0)
\end{align*}
\]

See Example 4.34.

Two-Parameter Weibull Distribution

To create the plot, the observations are ordered from smallest to largest, and the log of the shifted \( i \)th ordered observation \( x_{i(i)} \), denoted by \( \log(x_{i(i)} - \theta_0) \), is plotted against the quantile \( \log\left(-\log\left(1 - \frac{i-0.375}{n+0.25}\right)\right) \), where \( n \) is the number of nonmissing observations. In a probability plot, the horizontal axis is scaled in percentile units.

Unlike the three-parameter Weibull quantile, the preceding expression is free of distribution parameters. Consequently, the \( \text{C}= \) shape parameter is not mandatory with the \text{WEIBULL2} distribution option.

The pattern on the plot for \( \text{THETA}=\theta_0 \) 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

\[
\begin{align*}
\theta_0 &= \text{known lower threshold} \\
\sigma &= \text{scale parameter } (\sigma > 0) \\
c &= \text{shape parameter } (c > 0)
\end{align*}
\]

See Example 4.34.
### Estimating Shape Parameters Using Q-Q Plots

Some of the distribution options in the PROBPLOT or QQPLOT statements require you to specify one or two shape parameters in parentheses after the distribution keyword. These are summarized in Table 4.35.

You can visually estimate the value of a shape parameter by specifying a list of values for the shape parameter option. A separate plot is produced for each value, and you can then select the value of the shape parameter that produces the most nearly linear point pattern. Alternatively, you can request that the plot be created using an estimated shape parameter. See the entries for the distribution options in the section “Dictionary of Options” on page 350 (for the PROBPLOT statement) and in the section “Dictionary of Options” on page 361 (for the QQPLOT statement).

**NOTE:** For Q-Q plots created 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 $\frac{1}{c}$.

#### Table 4.35  Shape Parameter Options

<table>
<thead>
<tr>
<th>Distribution Keyword</th>
<th>Mandatory Shape Parameter Option</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETA</td>
<td>ALPHA=$\alpha$, BETA=$\beta$</td>
<td>$\alpha &gt; 0, \beta &gt; 0$</td>
</tr>
<tr>
<td>EXPONENTIAL</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>GAMMA</td>
<td>ALPHA=$\alpha$</td>
<td>$\alpha &gt; 0$</td>
</tr>
<tr>
<td>GUMBEL</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>LOGNORMAL</td>
<td>SIGMA=$\sigma$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td>NORMAL</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>PARETO</td>
<td>ALPHA=$\alpha$</td>
<td>$\alpha &gt; 0$</td>
</tr>
<tr>
<td>POWER</td>
<td>ALPHA=$\alpha$</td>
<td>$\alpha &gt; 0$</td>
</tr>
<tr>
<td>RAYLEIGH</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>WEIBULL</td>
<td>C=$c$</td>
<td>$c &gt; 0$</td>
</tr>
<tr>
<td>WEIBULL2</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>

### Estimating Location and Scale Parameters Using Q-Q Plots

If you specify location and scale parameters for a distribution in a PROBPLOT or QQPLOT statement (or if you request estimates for these parameters), a diagonal distribution reference line is displayed on the plot. (An exception is the two-parameter Weibull distribution, for which a line is displayed 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.

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.)

**Table 4.36** shows how the specified parameters determine the intercept and slope of the line. The intercept and slope are based on the quantile scale for the horizontal axis, which is used in Q-Q plots.
Table 4.36  Intercept and Slope of Distribution Reference Line

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameters</th>
<th>Linear Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta</td>
<td>$\theta$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\theta$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Gumbel</td>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\theta$</td>
<td>$\zeta$, $\sigma$</td>
</tr>
<tr>
<td>Normal</td>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Generalized Pareto</td>
<td>$\theta$</td>
<td>$\sigma$, $\alpha$</td>
</tr>
<tr>
<td>Power Function</td>
<td>$\theta$</td>
<td>$\sigma$, $\alpha$</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>$\theta$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Weibull (3-parameter)</td>
<td>$\theta$, $\sigma$, $c$</td>
<td>$\theta$</td>
</tr>
<tr>
<td>Weibull (2-parameter)</td>
<td>$\theta_0$ (known), $\sigma$, $c$</td>
<td>$\log(\sigma)$</td>
</tr>
</tbody>
</table>

For instance, specifying MU=3 and SIGMA=2 with the NORMAL option requests a line with intercept 3 and slope 2. Specifying SIGMA=1 and C=2 with the WEIBULL2 option requests a line with intercept $\log(1) = 0$ and slope $\frac{1}{2}$. On a probability plot with the LOGNORMAL and WEIBULL2 options, you can specify the slope directly with the SLOPE= option. That is, for the LOGNORMAL option, specifying THETA= $\theta_0$ and SLOPE=$\exp(\zeta_0)$ displays the same line as specifying THETA= $\theta_0$ and ZETA= $\zeta_0$. For the WEIBULL2 option, specifying SIGMA= $\sigma_0$ and SLOPE=$\frac{1}{c_0}$ displays the same line as specifying SIGMA= $\sigma_0$ and C= $c_0$.

Estimating Percentiles Using Q-Q Plots

There are two ways 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 50th percentile corresponds to the zero quantile.

You can also estimate percentiles by using probability plots created with the PROBPLOT statement. See Example 4.32.
Input Data Sets

The DATA= data set provides the set of variables that are analyzed. The UNIVARIATE procedure must have a DATA= data set. If you do not specify one with the DATA= option in the PROC UNIVARIATE statement, the procedure uses the last data set created.

ANNOTATE= Data Sets

You can add features to plots by specifying ANNOTATE= data sets either in the PROC UNIVARIATE statement or in individual plot statements.

Information contained in an ANNOTATE= data set specified in the PROC UNIVARIATE statement is used for all plots produced in a given PROC step; this is a “global” ANNOTATE= data set. By using this global data set, you can keep information common to all high-resolution plots in one data set.

Information contained in the ANNOTATE= data set specified in a plot statement is used only for plots produced by that statement; this is a “local” ANNOTATE= data set. By using this data set, you can add statement-specific features to plots. For example, you can add different features to plots produced by the HISTOGRAM and QQPLOT statements by specifying an ANNOTATE= data set in each plot statement.

You can specify an ANNOTATE= data set in the PROC UNIVARIATE statement and in plot statements. This enables you to add some features to all plots and also add statement-specific features to plots. See Example 4.25.

OUT= Output Data Set in the OUTPUT Statement

PROC UNIVARIATE creates an OUT= data set for each OUTPUT statement. This data set contains an observation for each combination of levels of the variables in the BY statement, or a single observation if you do not specify a BY statement. Thus the number of observations in the new data set corresponds to the number of groups for which statistics are calculated. Without a BY statement, the procedure computes statistics and percentiles by using all the observations in the input data set. With a BY statement, the procedure computes statistics and percentiles by using the observations within each BY group.

The variables in the OUT= data set are as follows:

- BY statement variables. The values of these variables match the values in the corresponding BY group in the DATA= data set and indicate which BY group each observation summarizes.

- variables created by selecting statistics in the OUTPUT statement. The statistics are computed using all the nonmissing data, or they are computed for each BY group if you use a BY statement.

- variables created by requesting new percentiles with the PCTLPTS= option. The names of these new variables depend on the values of the PCTLPRE= and PCTLNAME= options.

If the output data set contains a percentile variable or a quartile variable, the percentile definition assigned with the PCTLDDEF= option in the PROC UNIVARIATE statement is recorded in the output data set label. See Example 4.8.
The following table lists variables available in the OUT= data set.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>CSS</td>
<td>sum of squares corrected for the mean</td>
</tr>
<tr>
<td>CV</td>
<td>percent coefficient of variation</td>
</tr>
<tr>
<td>KURTOSIS</td>
<td>KURT</td>
</tr>
<tr>
<td>MAX</td>
<td>largest (maximum) value</td>
</tr>
<tr>
<td>MEAN</td>
<td>arithmetic mean</td>
</tr>
<tr>
<td>MIN</td>
<td>smallest (minimum) value</td>
</tr>
<tr>
<td>MODE</td>
<td>most frequent value (if not unique, the smallest mode)</td>
</tr>
<tr>
<td>N</td>
<td>number of observations on which calculations are based</td>
</tr>
<tr>
<td>NMISS</td>
<td>number of missing observations</td>
</tr>
<tr>
<td>NOBS</td>
<td>total number of observations</td>
</tr>
<tr>
<td>RANGE</td>
<td>difference between the maximum and minimum values</td>
</tr>
<tr>
<td>SKEWNESS</td>
<td>SKEW</td>
</tr>
<tr>
<td>STD</td>
<td>STDDEV</td>
</tr>
<tr>
<td>STDMEAN</td>
<td>STDERR</td>
</tr>
<tr>
<td>SUM</td>
<td>sum</td>
</tr>
<tr>
<td>SUMWGT</td>
<td>sum of the weights</td>
</tr>
<tr>
<td>USS</td>
<td>uncorrected sum of squares</td>
</tr>
<tr>
<td>VAR</td>
<td>variance</td>
</tr>
<tr>
<td><strong>Quantile Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>MEDIAN</td>
<td>Q2</td>
</tr>
<tr>
<td>P1</td>
<td>1st percentile</td>
</tr>
<tr>
<td>P5</td>
<td>5th percentile</td>
</tr>
<tr>
<td>P10</td>
<td>10th percentile</td>
</tr>
<tr>
<td>P90</td>
<td>90th percentile</td>
</tr>
<tr>
<td>P95</td>
<td>95th percentile</td>
</tr>
<tr>
<td>P99</td>
<td>99th percentile</td>
</tr>
<tr>
<td>Q1</td>
<td>P25</td>
</tr>
<tr>
<td>Q3</td>
<td>P75</td>
</tr>
<tr>
<td>QRANGE</td>
<td>difference between the upper and lower quartiles (also known as the inner quartile range)</td>
</tr>
<tr>
<td><strong>Robust Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>GINI</td>
<td>Gini’s mean difference</td>
</tr>
<tr>
<td>MAD</td>
<td>median absolute difference</td>
</tr>
<tr>
<td>QN</td>
<td>2nd variation of median absolute difference</td>
</tr>
<tr>
<td>SN</td>
<td>1st variation of median absolute difference</td>
</tr>
<tr>
<td>STD_GINI</td>
<td>standard deviation for Gini’s mean difference</td>
</tr>
<tr>
<td>STD_MAD</td>
<td>standard deviation for median absolute difference</td>
</tr>
<tr>
<td>STD_QN</td>
<td>standard deviation for the second variation of the median absolute difference</td>
</tr>
<tr>
<td>STD_QRANGE</td>
<td>estimate of the standard deviation, based on interquartile range</td>
</tr>
<tr>
<td>Variable Name</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>STD_SN</td>
<td>standard deviation for the first variation of the median absolute difference</td>
</tr>
<tr>
<td><strong>Hypothesis Test Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>MSIGN</td>
<td>sign statistic</td>
</tr>
<tr>
<td>NORMAL</td>
<td>test statistic for normality. If the sample size is less than or equal to 2000, this is the Shapiro-Wilk $W$ statistic. Otherwise, it is the Kolmogorov $D$ statistic.</td>
</tr>
<tr>
<td>PROBM</td>
<td>probability of a greater absolute value for the sign statistic</td>
</tr>
<tr>
<td>PROBN</td>
<td>probability that the data came from a normal distribution</td>
</tr>
<tr>
<td>PROBS</td>
<td>probability of a greater absolute value for the signed rank statistic</td>
</tr>
<tr>
<td>PROBT</td>
<td>two-tailed $p$-value for Student’s $t$ statistic with $n - 1$ degrees of freedom</td>
</tr>
<tr>
<td>SIGNRANK</td>
<td>signed rank statistic</td>
</tr>
<tr>
<td>T</td>
<td>Student’s $t$ statistic to test the null hypothesis that the population mean is equal to $\mu_0$</td>
</tr>
</tbody>
</table>

**OUTHISTOGRAM= Output Data Set**

You can create an OUTHISTOGRAM= data set with the HISTOGRAM statement. This data set contains information about histogram intervals. Because you can specify multiple HISTOGRAM statements with the UNIVARIATE procedure, you can create multiple OUTHISTOGRAM= data sets.

An OUTHISTOGRAM= data set contains a group of observations for each variable in the HISTOGRAM statement. 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 do not necessarily coincide with the intervals displayed in the histogram because the histogram might 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 an OUTHISTOGRAM= data set.

By default, an OUTHISTOGRAM= data set contains the _MIDPT_ variable, whose values identify histogram intervals by their midpoints. When the ENDPOINTS= or NENDPOINTS option is specified, intervals are identified by endpoint values instead. If the RTINCLUDE option is specified, the _MAXPT_ variable contains upper endpoint values. Otherwise, the _MINPT_ variable contains lower endpoint values. See Example 4.18.
Table 4.38 Variables in the OUTHISTOGRAM= Data Set

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>COUNT</em></td>
<td>number of variable values in histogram interval</td>
</tr>
<tr>
<td><em>CURVE</em></td>
<td>name of fitted distribution (if requested in HISTOGRAM statement)</td>
</tr>
<tr>
<td><em>EXPPCT</em></td>
<td>estimated percent of population in histogram interval determined from optional fitted distribution</td>
</tr>
<tr>
<td><em>MAXPT</em></td>
<td>upper endpoint of histogram interval</td>
</tr>
<tr>
<td><em>MIDPT</em></td>
<td>midpoint of histogram interval</td>
</tr>
<tr>
<td><em>MINPT</em></td>
<td>lower endpoint of histogram interval</td>
</tr>
<tr>
<td><em>OBSPCT</em></td>
<td>percent of variable values in histogram interval</td>
</tr>
<tr>
<td><em>VAR</em></td>
<td>variable name</td>
</tr>
</tbody>
</table>

OUTKERNEL= Output Data Set

You can create an OUTKERNEL= data set with the HISTOGRAM statement. This data set contains information about histogram intervals. Because you can specify multiple HISTOGRAM statements with the UNIVARIATE procedure, you can create multiple OUTKERNEL= data sets.

An OUTKERNEL= data set contains a group of observations for each kernel density estimate requested with the HISTOGRAM statement. These observations span a range of analysis variable values recorded in the _VALUE_ variable. The procedure determines the increment between values, and therefore the number of observations in the group. The variable _DENSITY_ contains the kernel density calculated for the corresponding analysis variable value.

When a density curve is overlaid on a histogram, the curve is scaled so that the area under the curve equals the total area of the histogram bars. The scaled density values are saved in the variable _COUNT_, _PERCENT_, or _PROPORTION_, depending on the histogram’s vertical axis scale, determined by the VSCALE= option. Only one of these variables appears in a given OUTKERNEL= data set.

Table 4.39 lists the variables in an OUTKERNEL= data set.

Table 4.39 Variables in the OUTKERNEL= Data Set

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C</em></td>
<td>standardized bandwidth parameter</td>
</tr>
<tr>
<td><em>COUNT</em></td>
<td>kernel density scaled for VSCALE=COUNT</td>
</tr>
<tr>
<td><em>DENSITY</em></td>
<td>kernel density</td>
</tr>
<tr>
<td><em>PERCENT</em></td>
<td>kernel density scaled for VSCALE=PERCENT (default)</td>
</tr>
<tr>
<td><em>PROPORTION</em></td>
<td>kernel density scaled for VSCALE=PROPORTION</td>
</tr>
<tr>
<td><em>TYPE</em></td>
<td>kernel function</td>
</tr>
<tr>
<td><em>VALUE</em></td>
<td>variable value at which kernel function is calculated</td>
</tr>
<tr>
<td><em>VAR</em></td>
<td>variable name</td>
</tr>
</tbody>
</table>
OUTTABLE= Output Data Set

The OUTTABLE= data set saves univariate statistics in a data set that contains one observation per analysis variable. The following variables are saved:

Table 4.40 Variables in the OUTTABLE= Data Set

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>CSS</em></td>
<td>corrected sum of squares</td>
</tr>
<tr>
<td><em>CV</em></td>
<td>coefficient of variation</td>
</tr>
<tr>
<td><em>GINI</em></td>
<td>Gini’s mean difference</td>
</tr>
<tr>
<td><em>KURT</em></td>
<td>kurtosis</td>
</tr>
<tr>
<td><em>MAD</em></td>
<td>median absolute difference about the median</td>
</tr>
<tr>
<td><em>MAX</em></td>
<td>maximum</td>
</tr>
<tr>
<td><em>MEAN</em></td>
<td>mean</td>
</tr>
<tr>
<td><em>MEDIAN</em></td>
<td>median</td>
</tr>
<tr>
<td><em>MIN</em></td>
<td>minimum</td>
</tr>
<tr>
<td><em>MODE</em></td>
<td>mode</td>
</tr>
<tr>
<td><em>MSIGN</em></td>
<td>sign statistic</td>
</tr>
<tr>
<td><em>NMISS</em></td>
<td>number of missing observations</td>
</tr>
<tr>
<td><em>NOBS</em></td>
<td>number of nonmissing observations</td>
</tr>
<tr>
<td><em>NORMAL</em></td>
<td>test statistic for normality</td>
</tr>
<tr>
<td><em>P1</em></td>
<td>1st percentile</td>
</tr>
<tr>
<td><em>P5</em></td>
<td>5th percentile</td>
</tr>
<tr>
<td><em>P10</em></td>
<td>10th percentile</td>
</tr>
<tr>
<td><em>P90</em></td>
<td>90th percentile</td>
</tr>
<tr>
<td><em>P95</em></td>
<td>95th percentile</td>
</tr>
<tr>
<td><em>P99</em></td>
<td>99th percentile</td>
</tr>
<tr>
<td><em>PROBM</em></td>
<td>p-value of sign statistic</td>
</tr>
<tr>
<td><em>PROBN</em></td>
<td>p-value of test for normality</td>
</tr>
<tr>
<td><em>PROBS</em></td>
<td>p-value of signed rank test</td>
</tr>
<tr>
<td><em>PROBT</em></td>
<td>p-value of t statistic</td>
</tr>
<tr>
<td><em>Q1</em></td>
<td>25th percentile (lower quartile)</td>
</tr>
<tr>
<td><em>Q3</em></td>
<td>75th percentile (upper quartile)</td>
</tr>
<tr>
<td><em>QN</em></td>
<td>$Q_n$</td>
</tr>
<tr>
<td><em>QRANGE</em></td>
<td>interquartile range (upper quartile minus lower quartile)</td>
</tr>
<tr>
<td><em>RANGE</em></td>
<td>range</td>
</tr>
<tr>
<td><em>SGNRNK</em></td>
<td>centered sign rank</td>
</tr>
<tr>
<td><em>SKEW</em></td>
<td>skewness</td>
</tr>
<tr>
<td><em>SN</em></td>
<td>$S_n$ (see “Robust Estimates of Scale” on page 388)</td>
</tr>
<tr>
<td><em>STD</em></td>
<td>standard deviation</td>
</tr>
<tr>
<td><em>STDGINI</em></td>
<td>Gini’s standard deviation</td>
</tr>
<tr>
<td><em>STDMAD</em></td>
<td>MAD standard deviation</td>
</tr>
<tr>
<td><em>STDMEAN</em></td>
<td>standard error of the mean</td>
</tr>
<tr>
<td><em>STDQN</em></td>
<td>$Q_n$ standard deviation</td>
</tr>
<tr>
<td><em>STDQRANGE</em></td>
<td>interquartile range standard deviation</td>
</tr>
<tr>
<td><em>STDSN</em></td>
<td>$S_n$ standard deviation</td>
</tr>
</tbody>
</table>
Table 4.40  (continued)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>SUMWGT</em></td>
<td>sum of the weights</td>
</tr>
<tr>
<td><em>SUM</em></td>
<td>sum</td>
</tr>
<tr>
<td><em>T</em></td>
<td>Student’s t statistic</td>
</tr>
<tr>
<td><em>USS</em></td>
<td>uncorrected sum of squares</td>
</tr>
<tr>
<td><em>VARI</em></td>
<td>variance</td>
</tr>
<tr>
<td><em>VAR</em></td>
<td>variable name</td>
</tr>
</tbody>
</table>

The OUTTABLE= data set and the OUT= data set see the section “OUT= Output Data Set in the OUTPUT Statement” on page 429 contain essentially the same information. in the OUT= data set (see the section “OUT= Output Data Set in the OUTPUT Statement” on page 429). However, the structure of the OUTTABLE= data set may be more appropriate when you are computing summary statistics for more than one analysis variable in the same invocation of the UNIVARIATE procedure. Each observation in the OUTTABLE= data set corresponds to a different analysis variable, and the variables in the data set correspond to summary statistics and indices.

For example, suppose you have 10 analysis variables (P1-P10). The following statements create an OUTTABLE= data set named Table, which contains summary statistics for each of these variables:

```sas
data Analysis;
  input A1-A10;
  datalines;
  72 223 332 138 110 145 23 293 353 458
  97 54 61 196 275 171 17 72 81 141
  56 170 140 400 371 72 60 20 484 138
  124 6 332 493 214 43 125 55 372 30
  152 236 222 76 187 126 192 334 109 546
  5 260 194 277 176 96 109 184 240 261
  161 253 153 300 37 156 282 293 451 299
  128 121 254 297 363 132 209 257 429 295
  116 152 331 27 442 103 80 393 383 94
  43 178 278 159 25 180 253 333 51 225
  34 128 182 415 524 112 13 186 145 131
  142 236 234 255 211 80 281 135 179 11
  108 215 335 66 254 196 190 363 226 379
  62 232 219 474 31 139 15 56 429 298
  177 218 275 171 457 146 163 18 155 129
  0 235 83 239 398 99 226 389 498 18
  147 199 324 258 504 2 218 295 422 287
  39 161 156 198 214 58 238 19 231 548
  120 42 372 420 232 112 157 79 197 166
  178 83 238 492 463 68 46 386 45 81
  161 267 372 296 501 96 11 288 330 74
  14 2 52 81 169 63 194 161 173 54
  22 181 92 272 417 94 188 180 367 342
  55 248 214 422 133 193 144 318 271 479
  56 83 169 30 379 5 296 320 396 597
;```
proc univariate data=Analysis outtable=Table noprint;
  var A1-A10;
run;

The following statements create the table shown in Figure 4.15, which contains the mean, standard deviation, and so on, for each analysis variable:

proc print data=Table label noobs;
  var _VAR_ _MIN_ _MEAN_ _MAX_ _STD_;
  label _VAR_='Analysis';
run;

Figure 4.15 Tabulating Results for Multiple Process Variables

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0</td>
<td>90.76</td>
<td>178</td>
<td>57.024</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td>167.32</td>
<td>267</td>
<td>81.628</td>
</tr>
<tr>
<td>A3</td>
<td>52</td>
<td>224.56</td>
<td>372</td>
<td>96.525</td>
</tr>
<tr>
<td>A4</td>
<td>27</td>
<td>258.08</td>
<td>493</td>
<td>145.218</td>
</tr>
<tr>
<td>A5</td>
<td>25</td>
<td>283.48</td>
<td>524</td>
<td>157.033</td>
</tr>
<tr>
<td>A6</td>
<td>2</td>
<td>107.48</td>
<td>196</td>
<td>52.437</td>
</tr>
<tr>
<td>A7</td>
<td>11</td>
<td>153.20</td>
<td>296</td>
<td>90.031</td>
</tr>
<tr>
<td>A8</td>
<td>18</td>
<td>217.08</td>
<td>393</td>
<td>130.031</td>
</tr>
<tr>
<td>A9</td>
<td>45</td>
<td>280.68</td>
<td>498</td>
<td>140.943</td>
</tr>
<tr>
<td>A10</td>
<td>11</td>
<td>243.24</td>
<td>597</td>
<td>178.799</td>
</tr>
</tbody>
</table>

Tables for Summary Statistics

By default, PROC UNIVARIATE produces ODS tables of moments, basic statistical measures, tests for location, quantiles, and extreme observations. You must specify options in the PROC UNIVARIATE statement to request other statistics and tables. The CIBASIC option produces a table that displays confidence limits for the mean, standard deviation, and variance. The CIPCTLDF and CIPCTLNORMAL options request tables of confidence limits for the quantiles. The LOCCOUNT option requests a table that shows the number of values greater than, not equal to, and less than the value of MU0=. The FREQ option requests a table of frequencies counts. The NEXTRVAL= option requests a table of extreme values. The NORMAL option requests a table with tests for normality.

The TRIMMED=, WINSORIZED=, and ROBUSTSCALE options request 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 $\sigma$.

See the section “ODS Table Names” on page 436 for the names of ODS tables created by PROC UNIVARIATE.
# ODS Table Names

PROC UNIVARIATE assigns a name to each table that it creates. You can use these names to reference the table when you use the Output Delivery System (ODS) to select tables and create output data sets.

**Table 4.41** ODS Tables Produced with the PROC UNIVARIATE Statement

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>BasicIntervals</td>
<td>confidence intervals for mean, standard deviation, variance</td>
<td>CIBASIC</td>
</tr>
<tr>
<td>BasicMeasures</td>
<td>measures of location and variability</td>
<td>default</td>
</tr>
<tr>
<td>ExtremeObs</td>
<td>extreme observations</td>
<td>default</td>
</tr>
<tr>
<td>ExtremeValues</td>
<td>extreme values</td>
<td>NEXTRVAL=</td>
</tr>
<tr>
<td>Frequencies</td>
<td>frequencies</td>
<td>FREQ</td>
</tr>
<tr>
<td>LocationCounts</td>
<td>counts used for sign test and signed rank test</td>
<td>LOCCOUNT</td>
</tr>
<tr>
<td>MissingValues</td>
<td>missing values</td>
<td>default, if missing values exist</td>
</tr>
<tr>
<td>Modes</td>
<td>modes</td>
<td>MODES</td>
</tr>
<tr>
<td>Moments</td>
<td>sample moments</td>
<td>default</td>
</tr>
<tr>
<td>Plots</td>
<td>line printer plots</td>
<td>PLOTS</td>
</tr>
<tr>
<td>Quantiles</td>
<td>quantiles</td>
<td>default</td>
</tr>
<tr>
<td>RobustScale</td>
<td>robust measures of scale</td>
<td>ROBUSTSCALE</td>
</tr>
<tr>
<td>SSPlots</td>
<td>line printer side-by-side box plots</td>
<td>PLOTS (with BY statement)</td>
</tr>
<tr>
<td>TestsForLocation</td>
<td>tests for location</td>
<td>default</td>
</tr>
<tr>
<td>TestsForNormality</td>
<td>tests for normality</td>
<td>NORMALTEST</td>
</tr>
<tr>
<td>TrimmedMeans</td>
<td>trimmed means</td>
<td>TRIMMED=</td>
</tr>
<tr>
<td>WinsorizedMeans</td>
<td>Winsorized means</td>
<td>WINSORIZED=</td>
</tr>
</tbody>
</table>

**Table 4.42** ODS Tables Produced with the HISTOGRAM Statement

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bins</td>
<td>histogram bins</td>
<td>MIDPERCENTS</td>
</tr>
<tr>
<td>FitQuantiles</td>
<td>quantiles of fitted distribution</td>
<td>secondary option</td>
</tr>
<tr>
<td>GoodnessOfFit</td>
<td>goodness-of-fit tests for fitted distribution</td>
<td>any distribution option</td>
</tr>
<tr>
<td>HistogramBins</td>
<td>histogram bins</td>
<td>MIDPERCENTS option</td>
</tr>
<tr>
<td>ParameterEstimates</td>
<td>parameter estimates for fitted distribution</td>
<td>any distribution option</td>
</tr>
</tbody>
</table>
ODS Tables for Fitted Distributions

If you request a fitted parametric distribution with a HISTOGRAM statement, PROC UNIVARIATE creates a summary that is organized into the ODS tables described in this section.

Parameters

The ParameterEstimates table lists the estimated (or specified) parameters for the fitted curve as well as the estimated mean and estimated standard deviation. See “Formulas for Fitted Continuous Distributions” on page 401.

EDF Goodness-of-Fit Tests

When you fit a parametric distribution, the HISTOGRAM statement provides a series of goodness-of-fit tests based on the empirical distribution function (EDF). See “EDF Goodness-of-Fit Tests” on page 414. These are displayed in the GoodnessOfFit table.

Histogram Intervals

The Bins table is included in the summary only if you specify the MIDPERCENTS option in parentheses after the distribution option. This table lists the midpoints for the histogram bins along with the observed and estimated percentages of the observations that lie in each bin. The estimated percentages are based on the fitted distribution.

If you specify the MIDPERCENTS option without requesting a fitted distribution, the HistogramBins table is included in the summary. This table lists the interval midpoints with the observed percent of observations that lie in the interval. See the entry for the MIDPERCENTS option on page 311.

Quantiles

The FitQuantiles table lists observed and estimated quantiles. You can use the PERCENTS= option to specify the list of quantiles in this table. See the entry for the PERCENTS= option on page 313. By default, the table lists observed and estimated quantiles for the 1, 5, 10, 25, 50, 75, 90, 95, and 99 percent of a fitted parametric distribution.

ODS Graphics


Before you create graphs, ODS Graphics must be enabled (for example, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” in that chapter.

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “A Primer on ODS Statistical Graphics” in that chapter.

PROC UNIVARIATE assigns a name to each graph it creates by using ODS Graphics. You can use these names to reference the graphs when you use ODS. The names are listed in Table 4.43.
Table 4.43  ODS Graphics Produced by PROC UNIVARIATE

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDFPlot</td>
<td>cdf plot</td>
<td>CDFPLOT</td>
</tr>
<tr>
<td>Histogram</td>
<td>histogram</td>
<td>HISTOGRAM</td>
</tr>
<tr>
<td>PPPlot</td>
<td>P-P plot</td>
<td>PPLOT</td>
</tr>
<tr>
<td>ProbPlot</td>
<td>probability plot</td>
<td>PROBPLOT</td>
</tr>
<tr>
<td>QQPlot</td>
<td>Q-Q plot</td>
<td>QQPLOT</td>
</tr>
</tbody>
</table>

Computational Resources

Because the UNIVARIATE procedure computes quantile statistics, it requires additional memory to store a copy of the data in memory. By default, the MEANS, SUMMARY, and TABULATE procedures 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-intensive.

In the UNIVARIATE procedure, 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 required 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 \( i \)th variable

Then the minimum memory requirement in bytes to process all variables is \( M = 24 \sum_i U_i \). 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 \( M = 24 \max(U_i) \).

Using the ROUND= option reduces the number of unique values \( U_i \), thereby reducing memory requirements. The ROBUSTSCALE option requires \( 40U_i \) bytes of temporary storage.

Several factors affect the CPU time:

- 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 \( i \)th 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 can 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 about optimizing CPU performance and memory usage, see the SAS documentation for your operating environment.
Example 4.1: Computing Descriptive Statistics for Multiple Variables

This example computes univariate statistics for two variables. The following statements create the data set BPressure, which contains the systolic (Systolic) and diastolic (Diastolic) blood pressure readings for 22 patients:

```sas
data BPressure;
  length PatientID $2;
  input PatientID $ Systolic  Diastolic @@;
datalines;
  CK  120  50  SS  96  60 FR  100  70
  CP  120  75  BL  140  90 ES  120  70
  CP  165 110  JI  110  40 MC  119  66
  FC  125  76  RW  133  60 KD  108  54
  DS 110  50  JW  130  80 BH  120  65
  JW 134  80  SB 118  76 NS  122  78
  GS 122  70  AB 122  78 EC  112  62
  HH 122  82;
```

The following statements produce descriptive statistics and quantiles for the variables Systolic and Diastolic:

```sas
title 'Systolic and Diastolic Blood Pressure';
ods select BasicMeasures Quantiles;
proc univariate data=BPressure;
  var Systolic Diastolic;
run;
```

The ODS SELECT statement restricts the output, which is shown in Output 4.1.1, to the “BasicMeasures” and “Quantiles” tables; see the section “ODS Table Names” on page 436. You use the PROC UNIVARIATE statement to request univariate statistics for the variables listed in the VAR statement, which specifies the analysis variables and their order in the output. Formulas for computing the statistics in the “BasicMeasures” table are provided in the section “Descriptive Statistics” on page 379. The quantiles are calculated using Definition 5, which is the default definition; see the section “Calculating Percentiles” on page 382.

A sample program for this example, uniex01.sas, is available in the SAS Sample Library for Base SAS software.
Output 4.1.1 Display Basic Measures and Quantiles

Systolic and Diastolic Blood Pressure

The UNIVARIATE Procedure
Variable: Systolic

Basic Statistical Measures

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>121.2727</td>
</tr>
<tr>
<td>Median</td>
<td>120.0000</td>
</tr>
<tr>
<td>Mode</td>
<td>120.0000</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>14.28346</td>
</tr>
<tr>
<td>Variance</td>
<td>204.01732</td>
</tr>
<tr>
<td>Range</td>
<td>69.00000</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>13.00000</td>
</tr>
</tbody>
</table>

Note: The mode displayed is the smallest of 2 modes with a count of 4.

Quantiles (Definition 5)

<table>
<thead>
<tr>
<th>Level</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>165</td>
</tr>
<tr>
<td>99%</td>
<td>165</td>
</tr>
<tr>
<td>95%</td>
<td>140</td>
</tr>
<tr>
<td>90%</td>
<td>134</td>
</tr>
<tr>
<td>75% Q3</td>
<td>125</td>
</tr>
<tr>
<td>50% Median</td>
<td>120</td>
</tr>
<tr>
<td>25% Q1</td>
<td>112</td>
</tr>
<tr>
<td>10%</td>
<td>108</td>
</tr>
<tr>
<td>5%</td>
<td>100</td>
</tr>
<tr>
<td>1%</td>
<td>96</td>
</tr>
<tr>
<td>0% Min</td>
<td>96</td>
</tr>
</tbody>
</table>

Systolic and Diastolic Blood Pressure

The UNIVARIATE Procedure
Variable: Diastolic

Basic Statistical Measures

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>70.09091</td>
</tr>
<tr>
<td>Median</td>
<td>70.00000</td>
</tr>
<tr>
<td>Mode</td>
<td>70.00000</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>15.16547</td>
</tr>
<tr>
<td>Variance</td>
<td>229.99134</td>
</tr>
<tr>
<td>Range</td>
<td>70.00000</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>18.00000</td>
</tr>
</tbody>
</table>

Example 4.2: Calculating Modes

An instructor is interested in calculating all the modes of the scores on a recent exam. The following statements create a data set named Exam, which contains the exam scores in the variable Score:

```plaintext
data Exam;
  label Score = 'Exam Score';
  input Score @@;
datalines;
81 97 78 99 77 81 84 86 86 97
85 86 94 76 75 42 91 90 88 86
97 97 89 69 72 82 83 81 80 81
;
```

The following statements use the MODES option to request a table of all possible modes:

```plaintext
title 'Table of Modes for Exam Scores';
ods select Modes;
proc univariate data=Exam modes;
  var Score;
run;
```

The ODS SELECT statement restricts the output to the “Modes” table; see the section “ODS Table Names” on page 436.
Output 4.2.1 Table of Modes Display

<table>
<thead>
<tr>
<th>Mode</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>4</td>
</tr>
<tr>
<td>86</td>
<td>4</td>
</tr>
<tr>
<td>97</td>
<td>4</td>
</tr>
</tbody>
</table>

By default, when the MODES option is used and there is more than one mode, the lowest mode is displayed in the “BasicMeasures” table. The following statements illustrate the default behavior:

```
title 'Default Output';
ods select BasicMeasures;
proc univariate data=Exam;
  var Score;
run;
```

Output 4.2.2 Default Output (Without MODES Option)

```
<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>83.66667</td>
</tr>
<tr>
<td>Median</td>
<td>84.50000</td>
</tr>
<tr>
<td>Mode</td>
<td>81.00000</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>11.08069</td>
</tr>
<tr>
<td>Variance</td>
<td>122.78161</td>
</tr>
<tr>
<td>Range</td>
<td>57.00000</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>10.00000</td>
</tr>
</tbody>
</table>

Note: The mode displayed is the smallest of 3 modes with a count of 4.
```

The default output displays a mode of 81 and includes a note regarding the number of modes; the modes 86 and 97 are not displayed. The ODS SELECT statement restricts the output to the “BasicMeasures” table; see the section “ODS Table Names” on page 436.

A sample program for this example, uniex02.sas, is available in the SAS Sample Library for Base SAS software.
Example 4.3: Identifying Extreme Observations and Extreme Values

This example, which uses the data set BPressure introduced in Example 4.1, illustrates how to produce a table of the extreme observations and a table of the extreme values in a data set. The following statements generate the “Extreme Observations” tables for Systolic and Diastolic, which enable you to identify the extreme observations for each variable:

```sas
title 'Extreme Blood Pressure Observations';
ods select ExtremeObs;
proc univariate data=BPressure;
   var Systolic Diastolic;
   id PatientID;
run;
```

The ODS SELECT statement restricts the output to the “ExtremeObs” table; see the section “ODS Table Names” on page 436. The ID statement requests that the extreme observations are to be identified using the value of PatientID as well as the observation number. By default, the five lowest and five highest observations are displayed. You can use the NEXTROBS= option to request a different number of extreme observations.

Output 4.3.1 shows that the patient identified as ‘CP’ (Observation 7) has the highest values for both Systolic and Diastolic. To visualize extreme observations, you can create histograms; see Example 4.14.

```
Output 4.3.1 Blood Pressure Extreme Observations

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>Patient</td>
<td>Obs</td>
<td>Value</td>
<td>Patient</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-----</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>96</td>
<td>SS</td>
<td>2</td>
<td>130</td>
<td>JW</td>
</tr>
<tr>
<td>100</td>
<td>FR</td>
<td>3</td>
<td>133</td>
<td>RW</td>
</tr>
<tr>
<td>108</td>
<td>KD</td>
<td>12</td>
<td>134</td>
<td>JW</td>
</tr>
<tr>
<td>110</td>
<td>DS</td>
<td>13</td>
<td>140</td>
<td>BL</td>
</tr>
<tr>
<td>110</td>
<td>JI</td>
<td>8</td>
<td>165</td>
<td>CP</td>
</tr>
</tbody>
</table>
```
The following statements generate the “Extreme Values” tables for Systolic and Diastolic, which tabulate the tails of the distributions:

```markdown
title 'Extreme Blood Pressure Values';
ods select ExtremeValues;
proc univariate data=BPressure nextrval=5;  
    var Systolic Diastolic;
run;
```

The ODS SELECT statement restricts the output to the “ExtremeValues” table; see the section “ODS Table Names” on page 436. The NEXTRVAL= option specifies the number of extreme values at each end of the distribution to be shown in the tables in Output 4.3.2.

Output 4.3.2 shows that the values 78 and 80 occurred twice for Diastolic and the maximum of Diastolic is 110. Note that Output 4.3.1 displays the value of 80 twice for Diastolic because there are two observations with that value. In Output 4.3.2, the value 80 is only displayed once.
Output 4.3.2  Blood Pressure Extreme Values

<table>
<thead>
<tr>
<th>Extreme Blood Pressure Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>The UNIVARIATE Procedure</td>
</tr>
<tr>
<td>Variable: Systolic</td>
</tr>
<tr>
<td>Extreme Values</td>
</tr>
<tr>
<td>------------------------------</td>
</tr>
<tr>
<td>Order</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extreme Blood Pressure Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>The UNIVARIATE Procedure</td>
</tr>
<tr>
<td>Variable: Diastolic</td>
</tr>
<tr>
<td>Extreme Values</td>
</tr>
<tr>
<td>------------------------------</td>
</tr>
<tr>
<td>Order</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

A sample program for this example, \textit{uniex01.sas}, is available in the SAS Sample Library for Base SAS software.

Example 4.4: Creating a Frequency Table

An instructor is interested in creating a frequency table of score changes between a pair of tests given in one of his college courses. The data set \textit{Score} contains test scores for his students who took a pretest and a posttest on the same material. The variable \textit{ScoreChange} contains the difference between the two test scores. The following statements create the data set:

```sas
data Score;
  input Student $ PreTest PostTest @@;
  label ScoreChange = 'Change in Test Scores';
  ScoreChange = PostTest - PreTest;
datalines;
```
The following statements produce a frequency table for the variable ScoreChange:

```sas
title 'Analysis of Score Changes';
ods select Frequencies;
proc univariate data=Score freq;
   var ScoreChange;
run;
```

The ODS SELECT statement restricts the output to the “Frequencies” table; see the section “ODS Table Names” on page 436. The FREQ option on the PROC UNIVARIATE statement requests the table of frequencies shown in Output 4.4.1.

**Output 4.4.1** Table of Frequencies

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percents</th>
</tr>
</thead>
<tbody>
<tr>
<td>-37</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>-14</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>-7</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>-5</td>
<td>2</td>
<td>16.7</td>
</tr>
<tr>
<td>-3</td>
<td>2</td>
<td>16.7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>100.0</td>
</tr>
</tbody>
</table>

From Output 4.4.1, the instructor sees that only score changes of $-3$ and $-5$ occurred more than once.

A sample program for this example, `uniex03.sas`, is available in the SAS Sample Library for Base SAS software.
Example 4.5: Creating Plots for Line Printer Output

When ODS Graphics is disabled, the PLOT option in the PROC UNIVARIATE statement requests several basic plots for display in line printer output. For more information about plots created by the PLOT option, see the section “Creating Line Printer Plots” on page 389. This example illustrates the use of the PLOT option as well as BY processing in PROC UNIVARIATE.

A researcher is analyzing a data set consisting of air pollution data from three different measurement sites. The data set AirPoll, created by the following statements, contains the variables Site and Ozone, which are the site number and ozone level, respectively.

```sas
data AirPoll (keep = Site Ozone);
  label Site = 'Site Number'
  Ozone = 'Ozone level (in ppb)';
  do i = 1 to 3;
    input Site @@;
    do j = 1 to 15;
      input Ozone @@;
      output;
    end;
  end;
  datalines;
102 4 6 3 4 7 8 2 3 4 1 3 8 9 5 6
134 5 3 6 2 1 2 4 3 2 4 6 4 6 3 1
137 8 9 7 8 6 7 6 7 9 8 9 8 7 8 5
;```

The following statements produce stem-and-leaf plots, box plots, and normal probability plots for each site in the AirPoll data set:

```sas
ods graphics off;
ods select Plots SSPlots;
proc univariate data=AirPoll plot;
  by Site;
  var Ozone;
run;
```

The PLOT option produces a stem-and-leaf plot, a box plot, and a normal probability plot for the Ozone variable at each site. Because the BY statement is used, a side-by-side box plot is also created to compare the ozone levels across sites. Note that AirPoll is sorted by Site; in general, the data set should be sorted by the BY variable by using the SORT procedure. The ODS SELECT statement restricts the output to the “Plots” and “SSPlots” tables; see the section “ODS Table Names” on page 436. Optionally, you can specify the PLOTSIZE=n option to control the approximate number of rows (between 8 and the page size) that the plots occupy.

Output 4.5.1 through Output 4.5.3 show the plots produced for each BY group. Output 4.5.4 shows the side-by-side box plot for comparing Ozone values across sites.
Output 4.5.1  Ozone Plots for BY Group Site = 102

Analysis of Score Changes
---------------------------------- Site Number=102 ----------------------------------

The UNIVARIATE Procedure
Variable: Ozone (Ozone level (in ppb))

Stem Leaf # Boxplot
9 0 1 | |
8 00 2 |
7 0 1 +------
6 00 2 | |
5 0 1 | |
4 000 3 *----*
3 000 3 +------
2 0 1 |
1 0 1 |

Normal Probability Plot
9.5+ *++++
| * * ++++
| *+++++
| *++++
5.5+ ++++
| ****
| * ++++
| ****
1.5+ *++++
+---------------------------------------------------------------
-2 -1 0 +1 +2
Output 4.5.2 Ozone Plots for BY Group Site = 134

Analysis of Score Changes

------------------------------- Site Number=134 --------------------------------

The UNIVARIATE Procedure
Variable: Ozone (Ozone level (in ppb))

<table>
<thead>
<tr>
<th>Stem Leaf</th>
<th>#</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 000</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5 0</td>
<td>1</td>
<td>++------</td>
</tr>
<tr>
<td>4 000</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3 000</td>
<td>3</td>
<td>*-------</td>
</tr>
<tr>
<td>2 000</td>
<td>3</td>
<td>+-------</td>
</tr>
<tr>
<td>1 00</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Normal Probability Plot

| 6.5+ | * * ++++++ |
|      | * ++++++ |
|      | * +++++ |
|      | **+++ |
|      | **++ |
| 1.5+ | * +++ |

-2 -1 0 +1 +2
Output 4.5.3  Ozone Plots for BY Group Site = 137

Analysis of Score Changes

------------------------------- Site Number=137 -----------------------------

The UNIVARIATE Procedure
Variable: Ozone (Ozone level (in ppb))

<table>
<thead>
<tr>
<th>Stem Leaf</th>
<th>#</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 000</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8 0000</td>
<td>5</td>
<td>+------</td>
</tr>
<tr>
<td>7 0000</td>
<td>4</td>
<td>+------</td>
</tr>
<tr>
<td>6 00</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5 0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Normal Probability Plot

9.5+     * ******
|     * * ******
7.5+     * * ******
|     ********
5.5+     ++++++  

-2 -1 0 +1 +2
Output 4.5.4 Ozone Side-by-Side Boxplot for All BY Groups

Analysis of Score Changes
The UNIVARIATE Procedure
Variable: Ozone (Ozone level (in ppb))

Schematic Plots

Analysis of Score Changes
The UNIVARIATE Procedure
Variable: Ozone (Ozone level (in ppb))

Schematic Plots

<table>
<thead>
<tr>
<th>Site</th>
<th>102</th>
<th>134</th>
<th>137</th>
</tr>
</thead>
</table>

Note that you can use the PROBPLOT statement with the NORMAL option to produce high-resolution normal probability plots; see the section “Modeling a Data Distribution” on page 271.

Note that you can use the BOXPLOT procedure to produce box plots that use high-resolution graphics. See Chapter 28, “The BOXPLOT Procedure” (SAS/STAT User’s Guide).

A sample program for this example, uniex04.sas, is available in the SAS Sample Library for Base SAS software.
Example 4.6: Analyzing a Data Set With a FREQ Variable

This example illustrates how to use PROC UNIVARIATE to analyze a data set with a variable that contains the frequency of each observation. The data set Speeding contains data on the number of cars pulled over for speeding on a stretch of highway with a 65 mile per hour speed limit. Speed is the speed at which the cars were traveling, and Number is the number of cars at each speed. The following statements create the data set:

```sas
DATA Speeding;
  LABEL Speed = 'Speed (in miles per hour)';
  DO Speed = 66 to 85;
    INPUT Number @@;
    OUTPUT;
  END;
DATAlines;
2 3 2 1 3 6 8 9 10 13
12 14 6 2 0 0 1 1 0 1;
```

The following statements create a table of moments for the variable Speed:

```sas
TITLE 'Analysis of Speeding Data';
ODS SELECT Moments;
PROC UNIVARIATE DATA=Speeding;
  FREQ Number;
  VAR Speed;
RUN;
```

The ODS SELECT statement restricts the output, which is shown in Output 4.6.1, to the “Moments” table; see the section “ODS Table Names” on page 436. The FREQ statement specifies that the value of the variable Number represents the frequency of each observation.

For the formulas used to compute these moments, see the section “Descriptive Statistics” on page 379. A sample program for this example, uniex05.sas, is available in the SAS Sample Library for Base SAS software.

**Output 4.6.1** Table of Moments

```
Analysis of Speeding Data

The UNIVARIATE Procedure
Variable:  Speed (Speed (in miles per hour))
Freq:  Number

Moments

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>94</td>
</tr>
<tr>
<td>Mean</td>
<td>74.3404255</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>3.44403237</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.1275543</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>520594</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>4.63278538</td>
</tr>
<tr>
<td>Sum Weights</td>
<td>6988</td>
</tr>
<tr>
<td>Sum Observations</td>
<td>11.861359</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.92002287</td>
</tr>
<tr>
<td>Corrected SS</td>
<td>1103.10638</td>
</tr>
<tr>
<td>Std Error Mean</td>
<td>0.35522482</td>
</tr>
</tbody>
</table>
```
Example 4.7: Saving Summary Statistics in an OUT= Output Data Set

This example illustrates how to save summary statistics in an output data set. The following statements create a data set named Belts, which contains the breaking strengths (Strength) and widths (Width) of a sample of 50 automotive seat belts:

```plaintext
data Belts;
  label Strength = 'Breaking Strength (lb/in)'
                 Width = 'Width in Inches';
  input Strength Width @@;
datalines;
1243.51  3.036  1221.95  2.995  1131.67  2.983  1129.70  3.019
1198.08  3.106  1273.31  2.947  1250.24  3.018  1225.47  2.980
1126.78  2.965  1174.62  3.033  1250.79  2.941  1216.75  3.037
1285.30  2.893  1214.14  3.035  1270.24  2.957  1249.55  2.958
1166.02  3.067  1278.85  3.037  1280.74  2.984  1201.96  3.002
1101.73  2.961  1165.79  3.075  1186.19  3.058  1124.46  2.929
1213.62  2.984  1213.93  3.029  1289.59  2.956  1208.27  3.029
1247.48  3.027  1284.34  3.073  1209.09  3.004  1146.78  3.061
1224.03  2.915  1200.43  2.974  1183.42  3.033  1195.66  2.995
1258.31  2.958  1136.05  3.022  1177.44  3.090  1246.13  3.022
1183.67  3.045  1206.50  3.024  1195.69  3.005  1223.49  2.971
1147.47  3.045  1171.76  3.005  1207.28  3.065  1131.33  2.984
1215.92  3.003  1202.17  3.058
;
```

The following statements produce two output data sets containing summary statistics:

```plaintext
proc univariate data=Belts noprint;
  var Strength Width;
  output out=Means     mean=StrengthMean WidthMean;
  output out=StrengthStats mean=StrengthMean std=StrengthSD
                 min=StrengthMin  max=StrengthMax;
run;
```

When you specify an OUTPUT statement, you must also specify a VAR statement. You can use multiple OUTPUT statements with a single procedure statement. Each OUTPUT statement creates a new data set with the name specified by the OUT= option. In this example, two data sets, Means and StrengthStats, are created. See Output 4.7.1 for a listing of Means and Output 4.7.2 for a listing of StrengthStats.

**Output 4.7.1** Listing of Output Data Set Means

<table>
<thead>
<tr>
<th>Analysis of Speeding Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 4: The UNIVARIATE Procedure

Output 4.7.2 Listing of Output Data Set StrengthStats

<table>
<thead>
<tr>
<th>Obs</th>
<th>Strength Mean</th>
<th>Strength SD</th>
<th>Strength Max</th>
<th>Strength Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1205.75</td>
<td>48.3290</td>
<td>1289.59</td>
<td>1101.73</td>
</tr>
</tbody>
</table>

Summary statistics are saved in an output data set by specifying `keyword=names` after the `OUT=` option. In the preceding statements, the first OUTPUT statement specifies the `keyword` MEAN followed by the `names` StrengthMean and WidthMean. The second OUTPUT statement specifies the `keywords` MEAN, STD, MAX, and MIN, for which the `names` StrengthMean, StrengthSD, StrengthMax, and StrengthMin are given.

The `keyword` specifies the statistic to be saved in the output data set, and the `names` determine the names for the new variables. The first `name` listed after a keyword contains that statistic for the first variable listed in the VAR statement; the second `name` contains that statistic for the second variable in the VAR statement, and so on.

The data set Means contains the mean of Strength in a variable named StrengthMean and the mean of Width in a variable named WidthMean. The data set StrengthStats contains the mean, standard deviation, maximum value, and minimum value of Strength in the variables StrengthMean, StrengthSD, StrengthMax, and StrengthMin, respectively.

See the section “OUT= Output Data Set in the OUTPUT Statement” on page 429 for more information about OUT= output data sets.

A sample program for this example, uniex06.sas, is available in the SAS Sample Library for Base SAS software.

Example 4.8: Saving Percentiles in an Output Data Set

This example, which uses the Belts data set from the previous example, illustrates how to save percentiles in an output data set. The UNIVARIATE procedure automatically computes the 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentiles for each variable. You can save these percentiles in an output data set by specifying the appropriate keywords. For example, the following statements create an output data set named PctlStrength, which contains the 5th and 95th percentiles of the variable Strength:

```sas
proc univariate data=Belts noprint;
  var Strength Width;
  output out=PctlStrength p5=p5str p95=p95str;
run;
```

The output data set PctlStrength is listed in Output 4.8.1.
Example 4.9: Computing Confidence Limits for the Mean, Standard Deviation, and Variance

You can use the PCTLPTS=, PCTLPRE=, and PCTLNAME= options to save percentiles not automatically computed by the UNIVARIATE procedure. For example, the following statements create an output data set named Pctls, which contains the 20th and 40th percentiles of the variables Strength and Width:

```sas
proc univariate data=Belts noprint;
var Strength Width;
output out=Pctls pctlpts = 20 40
      pctlpre = Strength Width
      pctlname = pct20 pct40;
run;
```

The PCTLPTS= option specifies the percentiles to compute (in this case, the 20th and 40th percentiles). The PCTLPRE= and PCTLNAME= options build the names for the variables containing the percentiles. The PCTLPRE= option gives prefixes for the new variables, and the PCTLNAME= option gives a suffix to add to the prefix. When you use the PCTLPTS= specification, you must also use the PCTLPRE= specification.

The OUTPUT statement saves the 20th and 40th percentiles of Strength and Width in the variables Strengthpct20, Widthpct20, Strengthpct40, and Weightpct40. The output data set Pctls is listed in Output 4.8.2.

Output 4.8.2 Listing of Output Data Set Pctls

```
Analysis of Speeding Data
Obs     Strengthpct20     Widthpct20     Strengthpct40     Weightpct40
1       1165.91           2.9595         1199.26           2.995
```

A sample program for this example, uniex06.sas, is available in the SAS Sample Library for Base SAS software.

Example 4.9: Computing Confidence Limits for the Mean, Standard Deviation, and Variance

This example illustrates how to compute confidence limits for the mean, standard deviation, and variance of a population. A researcher is studying the heights of a certain population of adult females. She has collected a random sample of heights of 75 females, which are saved in the data set Heights:
The following statements produce confidence limits for the mean, standard deviation, and variance of the population of heights:

```sas
data Heights;
  label Height = 'Height (in)';
  input Height @@;
datalines;
64.1 60.9 64.1 64.7 66.7 65.0 63.7 67.4 64.9 63.7
64.0 67.5 62.8 63.9 65.9 62.3 64.1 60.6 68.6 68.6
63.7 63.0 64.7 68.2 66.7 62.8 64.0 64.1 62.1 62.9
62.7 60.9 61.6 64.6 65.7 66.6 66.7 66.0 68.5 64.4
60.5 63.0 60.0 61.6 64.3 60.2 63.5 64.7 66.0 65.1
63.6 62.0 63.6 65.8 66.0 65.4 63.5 66.3 66.2 67.5
65.8 63.1 65.8 64.4 64.0 64.9 65.7 61.0 64.1 65.5
68.6 66.6 65.7 65.1 70.0
;
```

The CIBASIC option requests confidence limits for the mean, standard deviation, and variance. For example, Output 4.9.1 shows that the 95% confidence interval for the population mean is (64.06, 65.07). The ODS SELECT statement restricts the output to the “BasicIntervals” table; see the section “ODS Table Names” on page 436.

The confidence limits in Output 4.9.1 assume that the heights are normally distributed, so you should check this assumption before using these confidence limits. See the section “Shapiro-Wilk Statistic” on page 414 for information about the Shapiro-Wilk test for normality in PROC UNIVARIATE. See Example 4.19 for an example that uses the test for normality.

**Output 4.9.1** Default 95% Confidence Limits

<table>
<thead>
<tr>
<th>Analysis of Female Heights</th>
</tr>
</thead>
<tbody>
<tr>
<td>The UNIVARIATE Procedure</td>
</tr>
<tr>
<td>Variable: Height (Height (in))</td>
</tr>
</tbody>
</table>

<p>| Basic Confidence Limits Assuming Normality |
|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>64.56667</td>
<td>64.06302 65.07031</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>2.18900</td>
<td>1.88608 2.60874</td>
</tr>
<tr>
<td>Variance</td>
<td>4.79171</td>
<td>3.55731 6.80552</td>
</tr>
</tbody>
</table>

By default, the confidence limits produced by the CIBASIC option produce 95% confidence intervals. You can request different level confidence limits by using the ALPHA= option in parentheses after the CIBASIC option. The following statements produce 90% confidence limits:
Example 4.10: Computing Confidence Limits for Quantiles and Percentiles

This example, which is a continuation of Example 4.9, illustrates how to compute confidence limits for quantiles and percentiles. A second researcher is more interested in summarizing the heights with quantiles than the mean and standard deviation. He is also interested in computing 90% confidence intervals for the quantiles. The following statements produce estimated quantiles and confidence limits for the population quantiles:

```sas
title 'Analysis of Female Heights';
ods select Quantiles;
proc univariate data=Heights ciquantnormal(alpha=.1);
  var Height;
run;
```

The ODS SELECT statement restricts the output to the “Quantiles” table; see the section “ODS Table Names” on page 436. The CIQUANTNORMAL option produces confidence limits for the quantiles. As noted in Output 4.10.1, these limits assume that the data are normally distributed. You should check this assumption before using these confidence limits. See the section “Shapiro-Wilk Statistic” on page 414 for information about the Shapiro-Wilk test for normality in PROC UNIVARIATE; see Example 4.19 for an example that uses the test for normality.
Output 4.10.1  Normal-Based Quantile Confidence Limits

<table>
<thead>
<tr>
<th>Level</th>
<th>Quantile</th>
<th>Assuming Normality</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>70.0</td>
<td>70.0</td>
</tr>
<tr>
<td>99%</td>
<td>70.0</td>
<td>68.94553</td>
</tr>
<tr>
<td>95%</td>
<td>68.6</td>
<td>67.59184</td>
</tr>
<tr>
<td>90%</td>
<td>67.5</td>
<td>66.85981</td>
</tr>
<tr>
<td>75% Q3</td>
<td>66.0</td>
<td>65.60757</td>
</tr>
<tr>
<td>50% Median</td>
<td>64.4</td>
<td>64.14564</td>
</tr>
<tr>
<td>25% Q1</td>
<td>63.1</td>
<td>62.59071</td>
</tr>
<tr>
<td>10%</td>
<td>61.6</td>
<td>61.13060</td>
</tr>
<tr>
<td>5%</td>
<td>60.6</td>
<td>60.24022</td>
</tr>
<tr>
<td>1%</td>
<td>60.0</td>
<td>58.55106</td>
</tr>
<tr>
<td>0% Min</td>
<td>60.0</td>
<td></td>
</tr>
</tbody>
</table>

It is also possible to use PROC UNIVARIATE to compute confidence limits for quantiles without assuming normality. The following statements use the CIQUANTDF option to request distribution-free confidence limits for the quantiles of the population of heights:

```plaintext
title 'Analysis of Female Heights';
ods select Quantiles;
proc univariate data=Heights ciquantdf(alpha=.1);
    var Height;
run;
```

The distribution-free confidence limits are shown in Output 4.10.2.
Output 4.10.2 Distribution-Free Quantile Confidence Limits

Analysis of Female Heights

The UNIVARIATE Procedure
Variable: Height (Height (in))

Quantiles (Definition 5)

<table>
<thead>
<tr>
<th>Level</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>70.0</td>
</tr>
<tr>
<td>99%</td>
<td>70.0</td>
</tr>
<tr>
<td>95%</td>
<td>68.6</td>
</tr>
<tr>
<td>90%</td>
<td>67.5</td>
</tr>
<tr>
<td>75% Q3</td>
<td>66.0</td>
</tr>
<tr>
<td>50% Median</td>
<td>64.4</td>
</tr>
<tr>
<td>25% Q1</td>
<td>63.1</td>
</tr>
<tr>
<td>10%</td>
<td>61.6</td>
</tr>
<tr>
<td>5%</td>
<td>60.6</td>
</tr>
<tr>
<td>1%</td>
<td>60.0</td>
</tr>
<tr>
<td>0% Min</td>
<td>60.0</td>
</tr>
</tbody>
</table>

90% Confidence Limits

<table>
<thead>
<tr>
<th>Level</th>
<th>Distribution Free</th>
<th>LCL Rank</th>
<th>UCL Rank</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td></td>
<td>70.0</td>
<td>73</td>
<td>75</td>
</tr>
<tr>
<td>99%</td>
<td>68.6</td>
<td>70.0</td>
<td>68</td>
<td>75</td>
</tr>
<tr>
<td>95%</td>
<td>67.5</td>
<td>68.6</td>
<td>63</td>
<td>72</td>
</tr>
<tr>
<td>90% Q3</td>
<td>65.7</td>
<td>66.6</td>
<td>50</td>
<td>63</td>
</tr>
<tr>
<td>50% Median</td>
<td>64.1</td>
<td>65.1</td>
<td>31</td>
<td>46</td>
</tr>
<tr>
<td>25% Q1</td>
<td>62.7</td>
<td>63.7</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>10%</td>
<td>60.6</td>
<td>62.7</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>5%</td>
<td>60.0</td>
<td>61.6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>1%</td>
<td>60.0</td>
<td>60.5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0% Min</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table in Output 4.10.2 includes the ranks from which the confidence limits are computed. For more information about how these confidence limits are calculated, see the section “Confidence Limits for Percentiles” on page 383. Note that confidence limits for quantiles are not produced when the WEIGHT statement is used.

A sample program for this example, unixex07.sas, is available in the SAS Sample Library for Base SAS software.
Example 4.11: Computing Robust Estimates

This example illustrates how you can use the UNIVARIATE procedure to compute robust estimates of location and scale. The following statements compute these estimates for the variable Systolic in the data set BPressure, which was introduced in Example 4.1:

```plaintext
title 'Robust Estimates for Blood Pressure Data';
ods select TrimmedMeans WinsorizedMeans RobustScale;
proc univariate data=BPressure trimmed=1 .1
                             winsorized=.1 robustscale;
   var Systolic;
run;
```

The ODS SELECT statement restricts the output to the “TrimmedMeans,” “WinsorizedMeans,” and “RobustScale” tables; see the section “ODS Table Names” on page 436. The TRIMMED= option computes two trimmed means, the first after removing one observation and the second after removing 10% of the observations. If the value of TRIMMED= is greater than or equal to one, it is interpreted as the number of observations to be trimmed. The WINSORIZED= option computes a Winsorized mean that replaces three observations from the tails with the next closest observations. (Three observations are replaced because \( np = (22)(.1) = 2.2 \), and three is the smallest integer greater than 2.2.) The trimmed and Winsorized means for Systolic are displayed in Output 4.11.1.

**Output 4.11.1**  Computation of Trimmed and Winsorized Means

<table>
<thead>
<tr>
<th>Percent</th>
<th>Number</th>
<th>Mean</th>
<th>Std Error</th>
<th>95% Confidence Limits</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.55</td>
<td>1</td>
<td>120.3500</td>
<td>2.573536</td>
<td>114.9635 125.7365</td>
<td>19</td>
</tr>
<tr>
<td>13.64</td>
<td>3</td>
<td>120.3125</td>
<td>2.395387</td>
<td>115.2069 125.4181</td>
<td>15</td>
</tr>
</tbody>
</table>

**Trimmed Means**

| Percent | t for H0: Mu0=0.00 Pr > |t| |
|---------|-------------------------|-----|
| 4.55    | 46.76446                | <.0001 |
| 13.64   | 50.22675                | <.0001 |
Output 4.11.1 shows the trimmed mean for Systolic is 120.35 after one observation has been trimmed, and 120.31 after 3 observations are trimmed. The Winsorized mean for Systolic is 120.64. For details on trimmed and Winsorized means, see the section “Robust Estimators” on page 387. The trimmed means can be compared with the means shown in Output 4.1.1 (from Example 4.1), which displays the mean for Systolic as 121.273.

The ROBUSTSCALE option requests a table, displayed in Output 4.11.2, which includes the interquartile range, Gini’s mean difference, the median absolute deviation about the median, $Q_n$, and $S_n$.

Output 4.11.2 shows the robust estimates of scale for Systolic. For instance, the interquartile range is 13. The estimates of $\sigma$ range from 9.54 to 13.32. See the section “Robust Estimators” on page 387.

A sample program for this example, *unisex01.sas*, is available in the SAS Sample Library for Base SAS software.
Example 4.12: Testing for Location

This example, which is a continuation of Example 4.9, illustrates how to carry out three tests for location: the Student’s t test, the sign test, and the Wilcoxon signed rank test. These tests are discussed in the section “Tests for Location” on page 384.

The following statements demonstrate the tests for location by using the Heights data set introduced in Example 4.9. Because the data consists of adult female heights, the researchers are not interested in testing whether the mean of the population is equal to zero inches, which is the default $\mu_0$ value. Instead, they are interested in testing whether the mean is equal to 66 inches. The following statements test the null hypothesis $H_0: \mu_0 = 66$:

```sas
title 'Analysis of Female Height Data';
ods select TestsForLocation LocationCounts;
proc univariate data=Heights mu0=66 loccount;
   var Height;
run;
```

The ODS SELECT statement restricts the output to the “TestsForLocation” and “LocationCounts” tables; see the section “ODS Table Names” on page 436. The MU0= option specifies the null hypothesis value of $\mu_0$ for the tests for location; by default, $\mu_0 = 0$. The LOCCOUNT option produces the table of the number of observations greater than, not equal to, and less than 66 inches.

Output 4.12.1 contains the results of the tests for location. All three tests are highly significant, causing the researchers to reject the hypothesis that the mean is 66 inches.

A sample program for this example, `uniex07.sas`, is available in the SAS Sample Library for Base SAS software.

**Output 4.12.1** Tests for Location with MU0=66 and LOCCOUNT

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student’s t</td>
<td>t</td>
<td>Pr &gt;</td>
</tr>
<tr>
<td>Sign</td>
<td>M</td>
<td>Pr &gt;=</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S</td>
<td>Pr &gt;=</td>
</tr>
</tbody>
</table>

Location Counts: Mu0=66.00

<table>
<thead>
<tr>
<th>Count</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num Obs &gt; Mu0</td>
<td>16</td>
</tr>
<tr>
<td>Num Obs ^= Mu0</td>
<td>72</td>
</tr>
<tr>
<td>Num Obs &lt; Mu0</td>
<td>56</td>
</tr>
</tbody>
</table>
Example 4.13: Performing a Sign Test Using Paired Data

This example demonstrates a sign test for paired data, which is a specific application of the tests for location discussed in Example 4.12.

The instructor from Example 4.4 is now interested in performing a sign test for the pairs of test scores in his college course. The following statements request basic statistical measures and tests for location:

```sas
title 'Test Scores for a College Course';
ods select BasicMeasures TestsForLocation;
proc univariate data=Score;
    var ScoreChange;
run;
```

The ODS SELECT statement restricts the output to the “BasicMeasures” and “TestsForLocation” tables; see the section “ODS Table Names” on page 436. The instructor is not willing to assume that the `ScoreChange` variable is normal or even symmetric, so he decides to examine the sign test. The large \( p \)-value (0.7744) of the sign test provides insufficient evidence of a difference in test score medians.

Output 4.13.1 Sign Test for ScoreChange

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-3.08333</td>
</tr>
<tr>
<td>Median</td>
<td>-3.00000</td>
</tr>
<tr>
<td>Mode</td>
<td>-5.00000</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>13.33797</td>
</tr>
<tr>
<td>Variance</td>
<td>177.90152</td>
</tr>
<tr>
<td>Range</td>
<td>51.00000</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>10.50000</td>
</tr>
</tbody>
</table>

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

Tests for Location: \( \mu_0 = 0 \)

<table>
<thead>
<tr>
<th>Test</th>
<th>Test Statistic</th>
<th>( p )-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student's t</td>
<td>t = -0.80079</td>
<td>Pr &gt;</td>
</tr>
<tr>
<td>Sign</td>
<td>M = -1</td>
<td>Pr &gt;=</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S = -8.5</td>
<td>Pr &gt;=</td>
</tr>
</tbody>
</table>

A sample program for this example, `uniex03.sas`, is available in the SAS Sample Library for Base SAS software.
Chapter 4: The UNIVARIATE Procedure

Example 4.14: Creating a Histogram

This example illustrates how to create a histogram. A semiconductor manufacturer produces printed circuit boards that are sampled to determine the thickness of their copper plating. The following statements create a data set named `Trans`, which contains the plating thicknesses (`Thick`) of 100 boards:

```sas
data Trans;
  input Thick @@;
  label Thick = 'Plating Thickness (mils)';
  datalines;
; 
```

The following statements create the histogram shown in Output 4.14.1.

```sas
title 'Analysis of Plating Thickness';
ods graphics off;
proc univariate data=Trans noprint;
  histogram Thick;
run;
```

The NOPRINT option in the PROC UNIVARIATE statement suppresses tables of summary statistics for the variable `Thick` that would be displayed by default. A histogram is created for each variable listed in the HISTOGRAM statement.
Example 4.14: Creating a Histogram

Output 4.14.1  Histogram for Plating Thickness

A sample program for this example, *uniex08.sas*, is available in the SAS Sample Library for Base SAS software.
Example 4.15: Creating a One-Way Comparative Histogram

This example illustrates how to create a comparative histogram. The effective channel length (in microns) is measured for 1225 field effect transistors. The channel lengths (Length) are stored in a data set named Channel, which is partially listed in Output 4.15.1:

Output 4.15.1 Partial Listing of Data Set Channel

<table>
<thead>
<tr>
<th>Lot</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lot 1</td>
<td>0.91</td>
</tr>
<tr>
<td>Lot 1</td>
<td>1.17</td>
</tr>
<tr>
<td>Lot 2</td>
<td>1.47</td>
</tr>
<tr>
<td>Lot 2</td>
<td>1.39</td>
</tr>
<tr>
<td>Lot 3</td>
<td>2.04</td>
</tr>
<tr>
<td>Lot 3</td>
<td>1.91</td>
</tr>
</tbody>
</table>

The following statements request a histogram of Length ignoring the lot source:

```plaintext
   title 'Histogram of Length Ignoring Lot Source';
   ods graphics off;
   proc univariate data=Channel noprint;
      histogram Length;
   run;
```

The resulting histogram is shown in Output 4.15.2.
To investigate whether the peaks (modes) in Output 4.15.2 are related to the lot source, you can create a comparative histogram by using Lot as a classification variable. The following statements create the histogram shown in Output 4.15.3:

```sas
title 'Comparative Analysis of Lot Source';
ods graphics off;
proc univariate data=Channel noprint;
  class Lot;
  histogram Length / nrows = 3;
run;
```

The CLASS statement requests comparisons for each level (distinct value) of the classification variable Lot. The HISTOGRAM statement requests a comparative histogram for the variable Length. The NROWS= option specifies the number of rows per panel in the comparative histogram. By default, comparative histograms are displayed in two rows per panel.
Output 4.15.3 reveals that the distributions of Length are similarly distributed except for shifts in mean.

A sample program for this example, *uniex09.sas*, is available in the SAS Sample Library for Base SAS software.
Example 4.16: Creating a Two-Way Comparative Histogram

This example illustrates how to create a two-way comparative histogram. Two suppliers (A and B) provide disk drives for a computer manufacturer. The manufacturer measures the disk drive opening width to determine whether there has been a change in variability from 2002 to 2003 for each supplier.

The following statements save the measurements in a data set named Disk. There are two classification variables, Supplier and Year, and a user-defined format is associated with Year.

```sas
proc format;
    value mytime 1 = '2002' 2 = '2003';
data Disk;
    input @1 Supplier $10. Year Width;
    label Width = 'Opening Width (inches)';
    format Year mytime.;
datalines;
Supplier A 1 1.8932
         . .
Supplier B 1 1.8986
Supplier A 2 1.8978
         . .
Supplier B 2 1.8997
;
```

The following statements create the comparative histogram in Output 4.16.1:

```sas
title 'Results of Supplier Training Program';
ods graphics off;
proc univariate data=Disk noprint;
    class Supplier Year / keylevel = ('Supplier A' '2003');
    histogram Width / intertile = 1.0
        vaxis = 0 10 20 30
        ncols = 2
        nrows = 2;
run;
```

The KEYLEVEL= option specifies the key cell as the cell for which Supplier is equal to ‘SUPPLIER A’ and Year is equal to ‘2003.’ This cell determines the binning for the other cells, and the columns are arranged so that this cell is displayed in the upper left corner. Without the KEYLEVEL= option, the default key cell would be the cell for which Supplier is equal to ‘SUPPLIER A’ and Year is equal to ‘2002’; the column labeled ‘2002’ would be displayed to the left of the column labeled ‘2003.’

The VAXIS= option specifies the tick mark labels for the vertical axis. The NROWS=2 and NCOLS=2 options specify a $2 \times 2$ arrangement for the tiles. Output 4.16.1 provides evidence that both suppliers have reduced variability from 2002 to 2003.
A sample program for this example, `uniex10.sas`, is available in the SAS Sample Library for Base SAS software.
Example 4.17: Adding Insets with Descriptive Statistics

This example illustrates how to add insets with descriptive statistics to a comparative histogram; see Output 4.17.1. Three similar machines are used to attach a part to an assembly. One hundred assemblies are sampled from the output of each machine, and a part position is measured in millimeters. The following statements create the data set Machines, which contains the measurements in a variable named Position:

```sas
proc template;
  define StatGraph.HistBar;
    name = "Histogram with Inset Statistics";
    entry = "Histogram with Inset Statistics";
    data = "Machines";
    histvar = Position;
    histtitle = "Part Position in Millimeters";
    intertitle = "Interquartile Range";
    midrange = "Midpoint of the Bins";
    vaxis = "Value";
    haxis = "Category";
    plottype = "Bar"
  enddefine;
endproc;
```

```sas
data Machines;
  input Position @@;
  label Position = 'Position in Millimeters';
  if (_n_ <= 100) then Machine = 'Machine 1';
  else if (_n_ <= 200) then Machine = 'Machine 2';
  else Machine = 'Machine 3';
  datalines;
  -0.17 -0.19 -0.24 -0.24 -0.12 0.07 -0.61 0.22 1.91 -0.08
  -0.59 0.05 -0.38 0.82 -0.14 0.32 0.12 -0.02 0.26 0.19
  -0.07 0.13 -0.49 0.07 0.65 0.94 -0.51 -0.61 -0.57 -0.51
  ... more lines ...
  0.48 0.41 0.78 0.58 0.43 0.07 0.27 0.49 0.79 0.92
  0.79 0.66 0.22 0.71 0.53 0.57 0.90 0.48 1.17 1.03
;```

The following statements create the comparative histogram in Output 4.17.1:

```sas
title 'Machine Comparison Study';
ods graphics off;
proc univariate data=Machines noprint;
  class Machine;
  histogram Position / nrefs = 3
    intertitle = 1
    midpoints = -1.2 to 2.2 by 0.1
    vaxis = 0 to 16 by 4;
  inset mean std="Std Dev" / pos = ne format = 6.3;
run;
```

The INSET statement requests insets that contain the sample mean and standard deviation for each machine in the corresponding tile. The MIDPOINTS= option specifies the midpoints of the histogram bins.
Output 4.17.1 Comparative Histograms

Output 4.17.1 shows that the average position for Machines 2 and 3 are similar and that the spread for Machine 1 is much larger than for Machines 2 and 3.

A sample program for this example, uniex11.sas, is available in the SAS Sample Library for Base SAS software.
Example 4.18: Binning a Histogram

This example, which is a continuation of Example 4.14, demonstrates various methods for binning a histogram. This example also illustrates how to save bin percentages in an OUTHISTOGRAM= data set.

The manufacturer from Example 4.14 now wants to enhance the histogram by using the ENDPOINTS= option to change the endpoints of the bins. The following statements create a histogram with bins that have end points 3.425 and 3.6 and width 0.025:

```plaintext
title 'Enhancing a Histogram';
odds graphics off;
odds select HistogramBins MyHist;
proc univariate data=Trans;
   histogram Thick / midpercents name='MyHist'
               endpoints = 3.425 to 3.6 by .025;
run;
```

The ODS SELECT statement restricts the output to the “HistogramBins” table and the “MyHist” histogram; see the section “ODS Table Names” on page 436. The ENDPOINTS= option specifies the endpoints for the histogram bins. By default, if the ENDPOINTS= option is not specified, the automatic binning algorithm computes values for the midpoints of the bins. The MIDPERCENTS option requests a table of the midpoints of each histogram bin and the percent of the observations that fall in each bin. This table is displayed in Output 4.18.1; the histogram is displayed in Output 4.18.2. The NAME= option specifies a name for the histogram that can be used in the ODS SELECT statement.

**Output 4.18.1** Table of Bin Percentages Requested with MIDPERCENTS Option

<table>
<thead>
<tr>
<th>Bin Minimum Point</th>
<th>Observed Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.425</td>
<td>8.000</td>
</tr>
<tr>
<td>3.450</td>
<td>21.000</td>
</tr>
<tr>
<td>3.475</td>
<td>25.000</td>
</tr>
<tr>
<td>3.500</td>
<td>29.000</td>
</tr>
<tr>
<td>3.525</td>
<td>11.000</td>
</tr>
<tr>
<td>3.550</td>
<td>5.000</td>
</tr>
<tr>
<td>3.575</td>
<td>1.000</td>
</tr>
</tbody>
</table>
The MIDPOINTS= option is an alternative to the ENDPOINTS= option for specifying histogram bins. The following statements create a histogram, shown in Output 4.18.3, which is similar to the one in Output 4.18.2:

```plaintext
title 'Enhancing a Histogram';
ods graphics off;
proc univariate data=Trans noprint;
  histogram Thick / midpoints = 3.4375 to 3.5875 by .025
                        rttinclude
   outhistogram = OutMdpts;
run;
```

Output 4.18.3 differs from Output 4.18.2 in two ways:

- The MIDPOINTS= option specifies the bins for the histogram by specifying the midpoints of the bins instead of specifying the endpoints. Note that the histogram displays midpoints instead of endpoints.
- The RTINCLUDE option requests that the right endpoint of each bin be included in the histogram interval instead of the default, which is to include the left endpoint in the interval. This changes the histogram slightly from Output 4.18.2. Six observations have a thickness equal to an endpoint of an
interval. For instance, there is one observation with a thickness of 3.45 mils. In Output 4.18.3, this observation is included in the bin from 3.425 to 3.45.

Output 4.18.3 Histogram with MIDPOINTS= and RTINCLUDE Options

Output 4.18.4 The OUTHISTOGRAM= Data Set OutMdpts

The OUTHISTOGRAM= option produces an output data set named OutMdpts, displayed in Output 4.18.4. This data set provides information about the bins of the histogram. For more information, see the section “OUTHISTOGRAM= Output Data Set” on page 431.
A sample program for this example, uniex08.sas, is available in the SAS Sample Library for Base SAS software.

Example 4.19: Adding a Normal Curve to a Histogram

This example is a continuation of Example 4.14. The following statements fit a normal distribution to the thickness measurements in the Trans data set and superimpose the fitted density curve on the histogram:

```
title 'Analysis of Plating Thickness';
ods graphics off;
ods select ParameterEstimates GoodnessOfFit FitQuantiles Bins MyPlot;
proc univariate data=Trans;
    histogram Thick / normal(percents=20 40 60 80 midpercents)
         name='MyPlot';
    inset n normal(ksdpval) / pos = ne format = 6.3;
run;
```

The ODS SELECT statement restricts the output to the “ParameterEstimates,” “GoodnessOfFit,” “FitQuantiles,” and “Bins” tables; see the section “ODS Table Names” on page 436. The NORMAL option specifies that the normal curve be displayed on the histogram shown in Output 4.19.2. It also requests a summary of the fitted distribution, which is shown in Output 4.19.1. goodness-of-fit tests, parameter estimates, and quantiles of the fitted distribution. (If you specify the NORMALTEST option in the PROC UNIVARIATE statement, the Shapiro-Wilk test for normality is included in the tables of statistical output.)

Two secondary options are specified in parentheses after the NORMAL primary option. The PERCENTS= option specifies quantiles, which are to be displayed in the “FitQuantiles” table. The MIDPERCENTS option requests 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 normal distribution). See Table 4.6 for the secondary options that can be specified with after the NORMAL primary option.
Output 4.19.1 Summary of Fitted Normal Distribution

### Analysis of Plating Thickness

The UNIVARIATE Procedure

Fitted Normal Distribution for Thick (Plating Thickness (mils))

#### Parameters for Normal Distribution

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Mu</td>
<td>3.49533</td>
</tr>
<tr>
<td>Std Dev</td>
<td>Sigma</td>
<td>0.032117</td>
</tr>
</tbody>
</table>

#### Goodness-of-Fit Tests for Normal Distribution

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.05563823</td>
<td>&gt;0.150</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.04307548</td>
<td>&gt;0.250</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.27840748</td>
<td>&gt;0.250</td>
</tr>
</tbody>
</table>

#### Histogram Bin Percents for Normal Distribution

<table>
<thead>
<tr>
<th>Bin</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midpoint</td>
<td>Observed</td>
</tr>
<tr>
<td>3.43</td>
<td>3.000</td>
</tr>
<tr>
<td>3.45</td>
<td>9.000</td>
</tr>
<tr>
<td>3.47</td>
<td>23.000</td>
</tr>
<tr>
<td>3.49</td>
<td>19.000</td>
</tr>
<tr>
<td>3.51</td>
<td>24.000</td>
</tr>
<tr>
<td>3.53</td>
<td>15.000</td>
</tr>
<tr>
<td>3.55</td>
<td>3.000</td>
</tr>
<tr>
<td>3.57</td>
<td>4.000</td>
</tr>
</tbody>
</table>

#### Quantiles for Normal Distribution

<table>
<thead>
<tr>
<th>Percent</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>Estimated</td>
</tr>
<tr>
<td>20.0</td>
<td>3.46700</td>
</tr>
<tr>
<td>40.0</td>
<td>3.48350</td>
</tr>
<tr>
<td>60.0</td>
<td>3.50450</td>
</tr>
<tr>
<td>80.0</td>
<td>3.52250</td>
</tr>
</tbody>
</table>
The histogram of the variable Thick with a superimposed normal curve is shown in Output 4.19.2.

The estimated parameters for the normal curve ($\mu = 3.50$ and $\sigma = 0.03$) are shown in Output 4.19.1. By default, the parameters are estimated unless you specify values with the MU= and SIGMA= secondary options after the NORMAL primary option. The results of three goodness-of-fit tests based on the empirical distribution function (EDF) are displayed in Output 4.19.1. Because the $p$-values are all greater than 0.15, the hypothesis of normality is not rejected.

A sample program for this example, uniex08.sas, is available in the SAS Sample Library for Base SAS software.
This example is a continuation of Example 4.15, which introduced the data set Channel on page 466. In Output 4.15.3, it appears that the channel lengths in each lot are normally distributed. The following statements use the NORMAL option to fit a normal distribution for each lot:

```sas
title 'Comparative Analysis of Lot Source';
ods graphics off;
proc univariate data=Channel noprint;
   class Lot;
   histogram Length / nrows = 3
                   intertile = 1
                   cprop
                   normal(noprint);
   inset n = "N" / pos = nw;
run;
```

The NOPRINT option in the PROC UNIVARIATE statement suppresses the tables of statistical output produced by default; the NOPRINT option in parentheses after the NORMAL option suppresses the tables of statistical output related to the fit of the normal distribution. The normal parameters are estimated from the data for each lot, and the curves are superimposed on each component histogram. The INTERTILE= option specifies the space between the framed areas, which are referred to as tiles. The CPROP= option requests the shaded bars above each tile, which represent the relative frequencies of observations in each lot. The comparative histogram is displayed in Output 4.20.1.

A sample program for this example, `uniex09.sas`, is available in the SAS Sample Library for Base SAS software.
Output 4.20.1  Fitting Normal Curves to a Comparative Histogram

Comparative Analysis of Lot Source

<table>
<thead>
<tr>
<th>Lot</th>
<th>Percent</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lot 1</td>
<td></td>
<td>425</td>
</tr>
<tr>
<td>Lot 2</td>
<td></td>
<td>500</td>
</tr>
<tr>
<td>Lot 3</td>
<td></td>
<td>300</td>
</tr>
</tbody>
</table>

Length (microns)
Example 4.21: Fitting a Beta Curve

You can use a beta distribution to model the distribution of a variable that is known to vary between lower and upper bounds. In this example, a manufacturing company uses a robotic arm to attach hinges on metal sheets. The attachment point should be offset 10.1 mm from the left edge of the sheet. The actual offset varies between 10.0 and 10.5 mm due to variation in the arm. The following statements save the offsets for 50 attachment points as the values of the variable Length in the data set Robots:

```sas
data Robots;
   input Length @@;
   label Length = 'Attachment Point Offset (in mm)';
datalines;
10.147 10.070 10.032 10.042 10.102
10.034 10.143 10.278 10.114 10.127
10.122 10.018 10.271 10.293 10.136
10.240 10.205 10.186 10.186 10.080
10.158 10.114 10.018 10.201 10.065
10.122 10.139 10.090 10.136 10.066
10.074 10.175 10.052 10.059 10.077
10.211 10.122 10.031 10.322 10.187
10.094 10.067 10.094 10.051 10.174
;
```

The following statements create a histogram with a fitted beta density curve, shown in Output 4.21.1:

```sas
title 'Fitted Beta Distribution of Offsets';
ods graphics off;
ods select ParameterEstimates FitQuantiles MyHist;
proc univariate data=Robots;
   histogram Length /
      beta(theta=10 scale=0.5 color=red fill)
      href = 10
      hreflabel = 'Lower Bound'
      lhref = 2
      vaxis = axis1
      name = 'MyHist';
   axis1 label=(a=90 r=0);
   inset n = 'Sample Size'
      beta / pos=ne cfill=blank;
run;
```

The ODS SELECT statement restricts the output to the “ParameterEstimates” and “FitQuantiles” tables; see the section “ODS Table Names” on page 436. The BETA primary option requests a fitted beta distribution. The THETA= secondary option specifies the lower threshold. The SCALE= secondary option specifies the range between the lower threshold and the upper threshold. Note that the default THETA= and SCALE= values are zero and one, respectively.
The FILL secondary option specifies that the area under the curve is to be filled with the CFILL= color. (If FILL were omitted, the CFILL= color would be used to fill the histogram bars instead.)

The HREF= option draws a reference line at the lower bound, and the HREFLABEL= option adds the label Lower Bound. The LHREF= option specifies a dashed line type for the reference line. The INSET statement adds an inset with the sample size positioned in the northeast corner of the plot.

In addition to displaying the beta curve, the BETA option requests a summary of the curve fit. This summary, which includes parameters for the curve and the observed and estimated quantiles, is shown in Output 4.21.2.

A sample program for this example, uniex12.sas, is available in the SAS Sample Library for Base SAS software.
Example 4.22: Fitting Lognormal, Weibull, and Gamma Curves

To determine an appropriate model for a data distribution, you should consider curves from several distribution families. As shown in this example, you can use the HISTOGRAM statement to fit more than one distribution and display the density curves on a histogram.

The gap between two plates is measured (in cm) for each of 50 welded assemblies selected at random from the output of a welding process. The following statements save the measurements (Gap) in a data set named Plates:

data Plates;
  label Gap = 'Plate Gap in cm';
  input Gap @@;
  datalines;
0.746 0.357 0.376 0.327 0.485 1.741 0.241 0.777 0.768 0.409
0.252 0.512 0.534 1.656 0.742 0.378 0.714 1.121 0.597 0.231
0.541 0.805 0.682 0.418 0.506 0.501 0.247 0.922 0.880 0.344
0.519 1.302 0.275 0.601 0.388 0.450 0.845 0.319 0.486 0.529
1.547 0.690 0.676 0.314 0.736 0.643 0.483 0.352 0.636 1.080;

Output 4.21.2  Summary of Fitted Beta Distribution

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Theta</td>
<td>10</td>
</tr>
<tr>
<td>Scale</td>
<td>Sigma</td>
<td>0.5</td>
</tr>
<tr>
<td>Shape (Alpha)</td>
<td>Alpha</td>
<td>2.06832</td>
</tr>
<tr>
<td>Shape (Beta)</td>
<td>Beta</td>
<td>6.022479</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>10.12782</td>
</tr>
<tr>
<td>Std Dev</td>
<td></td>
<td>0.072339</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent</th>
<th>Observed</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>10.0180</td>
<td>10.0124</td>
</tr>
<tr>
<td>5.0</td>
<td>10.0310</td>
<td>10.0285</td>
</tr>
<tr>
<td>10.0</td>
<td>10.0380</td>
<td>10.0416</td>
</tr>
<tr>
<td>25.0</td>
<td>10.0670</td>
<td>10.0718</td>
</tr>
<tr>
<td>50.0</td>
<td>10.1220</td>
<td>10.1174</td>
</tr>
<tr>
<td>75.0</td>
<td>10.1750</td>
<td>10.1735</td>
</tr>
<tr>
<td>90.0</td>
<td>10.2255</td>
<td>10.2292</td>
</tr>
<tr>
<td>95.0</td>
<td>10.2780</td>
<td>10.2630</td>
</tr>
<tr>
<td>99.0</td>
<td>10.3220</td>
<td>10.3237</td>
</tr>
</tbody>
</table>
The following statements fit three distributions (lognormal, Weibull, and gamma) and display their density curves on a single histogram:

```plaintext
title 'Distribution of Plate Gaps';
ods graphics off;
ods select ParameterEstimates GoodnessOfFit FitQuantiles MyHist;
proc univariate data=Plates;
  var Gap;
  histogram / midpoints=0.2 to 1.8 by 0.2
           lognormal
           weibull
           gamma
           vaxis = axis1
           name = 'MyHist';
  inset n mean(5.3) std='Std Dev'(5.3) skewness(5.3)
                      / pos = ne header = 'Summary Statistics';
  axis1 label=(a=90 r=0);
run;
```

The ODS SELECT statement restricts the output to the “ParameterEstimates,” “GoodnessOfFit,” and “FitQuantiles” tables; see the section “ODS Table Names” on page 436. The LOGNORMAL, WEIBULL, and GAMMA primary options request superimposed fitted curves on the histogram in Output 4.22.1. Note that a threshold parameter $\theta = 0$ is assumed for each curve. In applications where the threshold is not zero, you can specify $\theta$ with the THETA= secondary option.

The LOGNORMAL, WEIBULL, and GAMMA options also produce the summaries for the fitted distributions shown in Output 4.22.2 through Output 4.22.4.

Output 4.22.2 provides three EDF goodness-of-fit tests for the lognormal distribution: the Anderson-Darling, the Cramér-von Mises, and the Kolmogorov-Smirnov tests. At the $\alpha = 0.10$ significance level, all tests support the conclusion that the two-parameter lognormal distribution with scale parameter $\xi = -0.58$ and shape parameter $\sigma = 0.50$ provides a good model for the distribution of plate gaps.
Example 4.22: Fitting Lognormal, Weibull, and Gamma Curves

Output 4.22.1 Superimposing a Histogram with Fitted Curves

Output 4.22.2 Summary of Fitted Lognormal Distribution

---

**Distribution of Plate Gaps**

The UNIVARIATE Procedure
Fitted Lognormal Distribution for Gap (Plate Gap in cm)

Parameters for Lognormal Distribution

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Theta</td>
<td>0</td>
</tr>
<tr>
<td>Scale</td>
<td>Zeta</td>
<td>-0.58375</td>
</tr>
<tr>
<td>Shape</td>
<td>Sigma</td>
<td>0.499546</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.631932</td>
</tr>
<tr>
<td>Std Dev</td>
<td></td>
<td>0.336436</td>
</tr>
</tbody>
</table>
### Output 4.22.2 continued

**Goodness-of-Fit Tests for Lognormal Distribution**

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov D</td>
<td>0.06441431</td>
<td>&gt;0.150</td>
</tr>
<tr>
<td>Cramer-von Mises W-Sq</td>
<td>0.02823022</td>
<td>&gt;0.500</td>
</tr>
<tr>
<td>Anderson-Darling A-Sq</td>
<td>0.24308402</td>
<td>&gt;0.500</td>
</tr>
</tbody>
</table>

**Quantiles for Lognormal Distribution**

<table>
<thead>
<tr>
<th>Percent</th>
<th>Observed</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.23100</td>
<td>0.17449</td>
</tr>
<tr>
<td>5.0</td>
<td>0.24700</td>
<td>0.24526</td>
</tr>
<tr>
<td>10.0</td>
<td>0.29450</td>
<td>0.29407</td>
</tr>
<tr>
<td>25.0</td>
<td>0.37800</td>
<td>0.39825</td>
</tr>
<tr>
<td>50.0</td>
<td>0.53150</td>
<td>0.55780</td>
</tr>
<tr>
<td>75.0</td>
<td>0.74600</td>
<td>0.78129</td>
</tr>
<tr>
<td>90.0</td>
<td>1.10050</td>
<td>1.05807</td>
</tr>
<tr>
<td>95.0</td>
<td>1.54700</td>
<td>1.26862</td>
</tr>
<tr>
<td>99.0</td>
<td>1.74100</td>
<td>1.78313</td>
</tr>
</tbody>
</table>

### Output 4.22.3 Summary of Fitted Weibull Distribution

**Distribution of Plate Gaps**

The UNIVARIATE Procedure
Fitted Weibull Distribution for Gap (Plate Gap in cm)

**Parameters for Weibull Distribution**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Theta</td>
<td>0</td>
</tr>
<tr>
<td>Scale</td>
<td>Sigma</td>
<td>0.719208</td>
</tr>
<tr>
<td>Shape</td>
<td>C</td>
<td>1.961159</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.637641</td>
</tr>
<tr>
<td>Std Dev</td>
<td></td>
<td>0.339248</td>
</tr>
</tbody>
</table>

**Goodness-of-Fit Tests for Weibull Distribution**

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cramer-von Mises W-Sq</td>
<td>0.15937281</td>
<td>0.016</td>
</tr>
<tr>
<td>Anderson-Darling A-Sq</td>
<td>1.15693542</td>
<td>&lt;0.010</td>
</tr>
</tbody>
</table>
**Output 4.22.3 continued**

<table>
<thead>
<tr>
<th>Percent</th>
<th>Observed</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.23100</td>
<td>0.06889</td>
</tr>
<tr>
<td>5.0</td>
<td>0.24700</td>
<td>0.15817</td>
</tr>
<tr>
<td>10.0</td>
<td>0.29450</td>
<td>0.22831</td>
</tr>
<tr>
<td>25.0</td>
<td>0.37800</td>
<td>0.38102</td>
</tr>
<tr>
<td>50.0</td>
<td>0.53150</td>
<td>0.59661</td>
</tr>
<tr>
<td>75.0</td>
<td>0.74600</td>
<td>0.84955</td>
</tr>
<tr>
<td>90.0</td>
<td>1.10050</td>
<td>1.10040</td>
</tr>
<tr>
<td>95.0</td>
<td>1.54700</td>
<td>1.25842</td>
</tr>
<tr>
<td>99.0</td>
<td>1.74100</td>
<td>1.56691</td>
</tr>
</tbody>
</table>

Output 4.22.3 provides two EDF goodness-of-fit tests for the Weibull distribution: the Anderson-Darling and the Cramér-von Mises tests. The \( p \)-values for the EDF tests are all less than 0.10, indicating that the data do not support a Weibull model.

**Output 4.22.4 Summary of Fitted Gamma Distribution**

<table>
<thead>
<tr>
<th>Distribution of Plate Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>The UNIVARIATE Procedure</td>
</tr>
<tr>
<td>Fitted Gamma Distribution for Gap (Plate Gap in cm)</td>
</tr>
</tbody>
</table>

**Parameters for Gamma Distribution**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Theta</td>
<td>0</td>
</tr>
<tr>
<td>Scale</td>
<td>Sigma</td>
<td>0.155198</td>
</tr>
<tr>
<td>Shape</td>
<td>Alpha</td>
<td>4.082646</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.63362</td>
</tr>
<tr>
<td>Std Dev</td>
<td></td>
<td>0.313587</td>
</tr>
</tbody>
</table>

**Goodness-of-Fit Tests for Gamma Distribution**

<table>
<thead>
<tr>
<th>Test</th>
<th>--------Statistic--------</th>
<th>------p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.09695325 Pr &gt; D</td>
<td>&gt;0.250</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.07398467 Pr &gt; W-Sq</td>
<td>&gt;0.250</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.58106613 Pr &gt; A-Sq</td>
<td>0.137</td>
</tr>
</tbody>
</table>
Output 4.22.4 provides three EDF goodness-of-fit tests for the gamma distribution: the Anderson-Darling, the Cramér–von Mises, and the Kolmogorov-Smirnov tests. At the $\alpha = 0.10$ significance level, all tests support the conclusion that the gamma distribution with scale parameter $\sigma = 0.16$ and shape parameter $\alpha = 4.08$ provides a good model for the distribution of plate gaps.

Based on this analysis, the fitted lognormal distribution and the fitted gamma distribution are both good models for the distribution of plate gaps.

A sample program for this example, uniex13.sas, is available in the SAS Sample Library for Base SAS software.

Example 4.23: Computing Kernel Density Estimates

This example illustrates the use of kernel density estimates to visualize a nonnormal data distribution. This example uses the data set Channel, which is introduced in Example 4.15.

When you compute kernel density estimates, you should try several choices for the bandwidth parameter $c$ because this determines the smoothness and closeness of the fit. You can specify a list of up to five C= values with the KERNEL option to request multiple density estimates, as shown in the following statements:

```sas
title 'FET Channel Length Analysis';
ods graphics off;
proc univariate data=Channel noprint;
   histogram Length / kernel(c = 0.25 0.50 0.75 1.00
         l = 1 20 2 34
         noprint);
run;
```

The L= secondary option specifies distinct line types for the curves (the L= values are paired with the C= values in the order listed). Output 4.23.1 demonstrates the effect of $c$. In general, larger values of $c$ yield smoother density estimates, and smaller values yield estimates that more closely fit the data distribution.
Example 4.24: Fitting a Three-Parameter Lognormal Curve

Output 4.23.1: Multiple Kernel Density Estimates

Output 4.23.1 reveals strong trimodality in the data, which is displayed with comparative histograms in Example 4.15.

A sample program for this example, *uniex09.sas*, is available in the SAS Sample Library for Base SAS software.

---

Example 4.24: Fitting a Three-Parameter Lognormal Curve

If you request a lognormal fit with the LOGNORMAL primary option, a two-parameter lognormal distribution is assumed. This means that the shape parameter \( \sigma \) and the scale parameter \( \xi \) are unknown (unless specified) and that the threshold \( \theta \) is known (it is either specified with the THETA= option or assumed to be zero).

If it is necessary to estimate \( \theta \) in addition to \( \xi \) and \( \sigma \), the distribution is referred to as a three-parameter lognormal distribution. This example shows how you can request a three-parameter lognormal distribution.

A manufacturing process produces a plastic laminate whose strength must exceed a minimum of 25 pounds per square inch (PSI). Samples are tested, and a lognormal distribution is observed for the strengths. It is important to estimate \( \theta \) to determine whether the process meets the strength requirement. The following statements save the strengths for 49 samples in the data set *Plastic*:
data Plastic;
  label Strength = 'Strength in psi';
  input Strength @@;
  datalines;
30.26 31.23 71.96 47.39 33.93 76.15 42.21
81.37 78.48 72.65 61.63 34.90 24.83 68.93
43.27 41.76 57.24 23.80 34.03 33.38 21.87
31.29 32.48 51.54 44.06 42.66 47.98 33.73
25.80 29.95 60.89 55.33 39.44 34.50 73.51
43.41 54.67 99.43 50.76 48.81 31.86 33.88
35.57 60.41 54.92 35.66 59.30 41.96 45.32
;

The following statements use the LOGNORMAL primary option in the HISTOGRAM statement to display the fitted three-parameter lognormal curve shown in Output 4.24.1:

```
title 'Three-Parameter Lognormal Fit';
ods graphics off;
proc univariate data=Plastic noprint;
  histogram Strength / lognormal(fill theta = est noprint);
  inset lognormal / format=6.2 pos=ne;
run;
```

The NOPRINT option suppresses the tables of statistical output produced by default. Specifying THETA=EST requests a local maximum likelihood estimate (LMLE) for $\theta$, as described by Cohen (1951). This estimate is then used to compute maximum likelihood estimates for $\sigma$ and $\xi$.

**NOTE:** You can also specify THETA=EST with the WEIBULL primary option to fit a three-parameter Weibull distribution.

A sample program for this example, `uniex14.sas`, is available in the SAS Sample Library for Base SAS software.
Example 4.25: Annotating a Folded Normal Curve

This example shows how to display a fitted curve that is not supported by the HISTOGRAM statement. The offset of an attachment point is measured (in mm) for a number of manufactured assemblies, and the measurements (Offset) are saved in a data set named Assembly. The following statements create the data set Assembly:

```plaintext
data Assembly;
  label Offset = 'Offset (in mm)';
  input Offset @@;
datalines;
11.11 13.07 11.42 3.92 11.08 5.40 11.22 14.69 6.27 9.76
9.97 3.28 13.03 13.78 3.13 9.53 4.58 7.94 13.51 11.43
11.98 3.90 7.67 4.32 12.69 6.17 11.48 2.82 20.42 1.01
3.18 6.02 6.63 1.72 2.42 11.32 16.49 1.22 9.13 3.34
1.29 1.70 0.65 2.62 2.04 11.08 18.85 11.94 8.34 2.07
0.31 8.91 13.62 14.94 4.83 16.84 7.09 3.37 0.49 15.19
5.16 4.14 1.92 12.70 1.97 2.10 9.38 3.18 4.18 7.22
```

Output 4.24.1 Three-Parameter Lognormal Fit

![Three-Parameter Lognormal Fit Graph]

- Lognormal -
It is decided to fit a folded normal distribution to the offset measurements. A variable $X$ has a folded normal distribution if $X = \lvert Y \rvert$, where $Y$ is distributed as $N(\mu, \sigma)$. The fitted density is

$$h(x) = \frac{1}{\sqrt{2\pi} \sigma} \left[ \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) + \exp \left( -\frac{(x + \mu)^2}{2\sigma^2} \right) \right]$$

where $x \geq 0$.

You can use SAS/IML to compute preliminary estimates of $\mu$ and $\sigma$ based on a method of moments given by Elandt (1961). These estimates are computed by solving equation (19) Elandt (1961), which is given by

$$f(\theta) = \frac{\left( \frac{2}{2\sqrt{\pi}} e^{-\theta^2/2} - \theta [1 - 2\Phi(\theta)] \right)^2}{1 + \theta^2} = A$$

where $\Phi(\cdot)$ is the standard normal distribution function and

$$A = \frac{\hat{x}^2}{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$$

Then the estimates of $\sigma$ and $\mu$ are given by

$$\hat{\sigma}_0 = \sqrt{\frac{\frac{1}{n} \sum_{i=1}^{n} x_i^2}{1 + \theta^2}}$$

$$\hat{\mu}_0 = \hat{\theta} \cdot \hat{\sigma}_0$$

Begin by using PROC MEANS to compute the first and second moments and by using the following DATA step to compute the constant $A$:

```sas
proc means data = Assembly noprint;
   var Offset;
   output out=stat mean=m1 var=var n=n min = min;
run;

* Compute constant A from equation (19) of Elandt (1961);
data stat;
   keep m2 a min;
   set stat;
   a = (m1*m1);
   m2 = ((n-1)/n)*var + a;
   a = a/m2;
run;
```

Next, use the SAS/IML subroutine NLPDD to solve equation (19) by minimizing $(f(\theta) - A)^2$, and compute $\hat{\mu}_0$ and $\hat{\sigma}_0$: 

```
15.84 10.85 2.35 1.93 9.19 1.39 11.40 12.20 16.07 9.23
0.05 2.15 1.95 4.39 0.48 10.16 4.81 8.28 5.68 22.81
0.23 0.38 12.71 0.06 10.11 18.38 5.53 9.36 9.32 3.63
12.93 10.39 2.05 15.49 8.12 9.52 7.77 10.70 6.37 1.91
8.60 22.22 1.74 5.84 12.90 13.06 5.08 2.09 6.41 1.40
12.91 0.95 0.89 3.82 7.86 5.33 12.92 2.64 7.92 14.06
```
Example 4.25: Annotating a Folded Normal Curve

```plaintext
proc iml;
    use stat;
    read all var {m2} into m2;
    read all var {a} into a;
    read all var {min} into min;

    * f(t) is the function in equation (19) of Elandt (1961);
    start f(t) global(a);
        y = .39894*exp(-0.5*t*t);
        y = (2*y-(t*(1-2*probnorm(t))))**2/(1+t*t);
        y = (y-a)**2;
        return(y);
    finish;

    * Minimize (f(t)-A)**2 and estimate mu and sigma;
    if ( min < 0 ) then do;
        print "Warning: Observations are not all nonnegative."
        print " The folded normal is inappropriate."
        stop;
    end;
    if ( a < 0.637 ) then do;
        print "Warning: the folded normal may be inappropriate"
    end;
    opt = { 0 0 };  
    con = { 1e-6 };
    x0 = { 2.0 };
    tc = { . . . . . 1e-8 . . . . . };
    call nlpdd(rc,etheta0,"f",x0,opt,con,tc);
    esig0 = sqrt(m2/(1+etheta0*etheta0));
    emu0 = etheta0*esig0;

    create prelim var {emu0 esig0 etheta0};
    append;
    close prelim;

    * Define the log likelihood of the folded normal;
    start g(p) global(x);
        y = 0.0;
        do i = 1 to nrow(x);
            z = exp( -0.5/p[2])*(x[i]-p[1])*(x[i]-p[1] );
            z = z + exp( -0.5/p[2])*(x[i]+p[1])*(x[i]+p[1] );
            y = y + log(z);
        end;
        y = y - nrow(x)*log( sqrt( p[2] ) );
        return(y);
    finish;

    * Maximize the log likelihood with subroutine NLPDD;
    use assembly;
    read all var {offset} into x;
    esig0sq = esig0*esig0;
    x0 = emu0||esig0sq;
    opt = { 1 0 };
    con = { . 0.0, . . };
```

The preliminary estimates are saved in the data set Prelim, as shown in Output 4.25.1.

### Output 4.25.1 Preliminary Estimates of μ, σ, and θ

<table>
<thead>
<tr>
<th>Obs</th>
<th>EMU0</th>
<th>ESIG0</th>
<th>ETHETA0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.51735</td>
<td>6.54953</td>
<td>0.99509</td>
</tr>
</tbody>
</table>

Now, using $\hat{\mu}_0$ and $\hat{\sigma}_0$ as initial estimates, call the NLPDD subroutine to maximize the log likelihood, $l(\mu, \sigma)$, of the folded normal distribution, where, up to a constant,

$$l(\mu, \sigma) = -n \log \sigma + \sum_{i=1}^{n} \left[ \exp \left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right) + \exp \left( -\frac{(x_i + \mu)^2}{2\sigma^2} \right) \right]$$

* Define the log likelihood of the folded normal;

```plaintext
start g(p) global(x);
    y = 0.0;
    do i = 1 to nrow(x);
        z = exp( (-0.5/p[2])*(x[i]-p[1])*(x[i]-p[1]) );
        z = z + exp( (-0.5/p[2])*(x[i]+p[1])*(x[i]+p[1]) );
        y = y + log(z);
    end;
    y = y - nrow(x)*log( sqrt( p[2] ) );
    return(y);
finish;
```

* Maximize the log likelihood with subroutine NLPDD;

```plaintext
use assembly;
read all var {offset} into x;
esig0sq = esig0*esig0;
x0 = emu0||esig0sq;
opt = { 1 0 };
con = { . 0.0, . . };
call nlpdd(rc,xr,"g",x0,opt,con);
emu = xr[1];
esig = sqrt(xr[2]);
etheta = emu/esig;
create parmest var{emu esig etheta};
append;
close parmest;
quit;
```

The data set ParmEst contains the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ (as well as $\hat{\mu}/\hat{\sigma}$), as shown in Output 4.25.2.
Example 4.25: Annotating a Folded Normal Curve

Output 4.25.2 Final Estimates of $\mu$, $\sigma$, and $\theta$

<table>
<thead>
<tr>
<th>Obs</th>
<th>EMU</th>
<th>ESIG</th>
<th>ETHETA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.66761</td>
<td>6.39650</td>
<td>1.04239</td>
</tr>
</tbody>
</table>

To annotate the curve on a histogram, begin by computing the width and endpoints of the histogram intervals. The following statements save these values in a data set called OutCalc. Note that a plot is not produced at this point.

```sas
proc univariate data = Assembly noprint;
   histogram Offset / outhistogram = out normal(noprint) noplot;
run;

data OutCalc (drop = _MIDPT_);
   set out (keep = _MIDPT_) end = eof;
   retain _MIDPT1_ _WIDTH_;  
   if _N_ = 1 then _MIDPT1_ = _MIDPT_; 
   if eof then do;
      _MIDPTN_ = _MIDPT_; 
      _WIDTH_ = (_MIDPTN_ - _MIDPT1_) / (_N_ - 1); 
      output;
   end;
run;
```

Output 4.25.3 provides a listing of the data set OutCalc. The width of the histogram bars is saved as the value of the variable _WIDTH_; the midpoints of the first and last histogram bars are saved as the values of the variables _MIDPT1_ and _MIDPTN_.

**Output 4.25.3** The Data Set OutCalc

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>MIDPT1</em></th>
<th><em>WIDTH</em></th>
<th><em>MIDPTN</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
<td>3</td>
<td>22.5</td>
</tr>
</tbody>
</table>

The following statements create an annotate data set named Anno, which contains the coordinates of the fitted curve:

```sas
data Anno;
   merge ParmEst OutCalc;
   length function color $ 8;
   function = 'point';
   color = 'black';
   size = 2;
   xsys = '2';
   ysys = '2';
   when = 'a';
   constant = 39.894*_width_;  
   left = _midpt1_ - .5*_width_; 
   right = _midptn_ + .5*_width_;  
run;
```
The following statements read the ANNOTATE= data set and display the histogram and fitted curve:

```plaintext
title 'Folded Normal Distribution';
ods graphics off;
proc univariate data=assembly noprint;
   histogram Offset / annotate = anno;
run;
```

Output 4.25.4 displays the histogram and fitted curve.
A sample program for this example, *uniex15.sas*, is available in the SAS Sample Library for Base SAS software.

---

**Example 4.26: Creating Lognormal Probability Plots**

This example is a continuation of the example explored in the section “Modeling a Data Distribution” on page 271.

In the normal probability plot shown in Output 4.6, the nonlinearity of the point pattern indicates a departure from normality in the distribution of Deviation. Because the point pattern is curved with slope increasing from left to right, a theoretical distribution that is skewed to the right, such as a lognormal distribution, should provide a better fit than the normal distribution. See the section “Interpretation of Quantile-Quantile and Probability Plots” on page 420.

You can explore the possibility of a lognormal fit with a lognormal probability plot. When you request such a plot, you must specify the shape parameter \( \sigma \) for the lognormal distribution. This value must be positive, and typical values of \( \sigma \) range from 0.1 to 1.0. You can specify values for \( \sigma \) with the SIGMA= secondary option in the LOGNORMAL primary option, or you can specify that \( \sigma \) is to be estimated from the data.

The following statements illustrate the first approach by creating a series of three lognormal probability plots for the variable Deviation introduced in the section “Modeling a Data Distribution” on page 271:

```sas
symbol v=plus height=3.5pct;
title 'Lognormal Probability Plot for Position Deviations';
ods graphics off;
proc univariate data=Aircraft noprint;
    probplot Deviation / lognormal(theta=est zeta=est sigma=0.7 0.9 1.1) href = 95 lhref = 1 square;
run;
```

The LOGNORMAL primary option requests plots based on the lognormal family of distributions, and the SIGMA= secondary option requests plots for \( \sigma \) equal to 0.7, 0.9, and 1.1. These plots are displayed in Output 4.26.1, Output 4.26.2, and Output 4.26.3, respectively. Alternatively, you can specify \( \sigma \) to be estimated using the sample standard deviation by using the option SIGMA=EST.

The SQUARE option displays the probability plot in a square format, the HREF= option requests a reference line at the 95th percentile, and the LHREF= option specifies the line type for the reference line.
Output 4.26.1  Probability Plot Based on Lognormal Distribution with $\sigma = 0.7$
Output 4.26.2 Probability Plot Based on Lognormal Distribution with $\sigma = 0.9$
The value $\sigma = 0.9$ in Output 4.26.2 most nearly linearizes the point pattern. The 95th percentile of the position deviation distribution seen in Output 4.26.2 is approximately 0.001, because this is the value corresponding to the intersection of the point pattern with the reference line.

**NOTE:** After the $\sigma$ that produces the most linear fit is found, you can then estimate the threshold parameter $\theta$ and the scale parameter $\zeta$. See Example 4.31.

The following statements illustrate how you can create a lognormal probability plot for Deviation by using a local maximum likelihood estimate for $\sigma$.

```plaintext
symbol v=plus height=3.5pct;
title 'Lognormal Probability Plot for Position Deviations';
ods graphics off;
proc univariate data=Aircraft noprint;
   probplot Deviation / lognormal(theta=est zeta=est sigma=est)
      href = 95
      square;
run;
```

The plot is displayed in Output 4.26.4. Note that the maximum likelihood estimate of $\sigma$ (in this case, 0.882) does not necessarily produce the most linear point pattern.
Example 4.27: Creating a Histogram to Display Lognormal Fit

This example uses the data set Aircraft from Example 4.26 to illustrate how to display a lognormal fit with a histogram. To determine whether the lognormal distribution is an appropriate model for a distribution, you should consider the graphical fit as well as conduct goodness-of-fit tests. The following statements fit a lognormal distribution and display the density curve on a histogram:

```sas
title 'Distribution of Position Deviations';
ods graphics off;
proc univariate data=Aircraft;
  var Deviation;
  histogram / lognormal(w=3 theta=est)
    vaxis = axis1
    name = 'MyPlot';
```

A sample program for this example, *unixex16.sas*, is available in the SAS Sample Library for Base SAS software.
The ODS SELECT statement restricts the output to the “ParameterEstimates” and “GoodnessOfFit” tables; see the section “ODS Table Names” on page 436. The LOGNORMAL primary option superimposes a fitted curve on the histogram in Output 4.27.1. The W= option specifies the line width for the curve. The INSET statement specifies that the mean, standard deviation, and skewness be displayed in an inset in the northeast corner of the plot. Note that the default value of the threshold parameter \( \theta \) is zero. In applications where the threshold is not zero, you can specify \( \theta \) with the THETA= option. The variable Deviation includes values that are less than the default threshold; therefore, the option THETA= EST is used.

Output 4.27.1 Normal Probability Plot Created with Graphics Device

Output 4.27.2 provides three EDF goodness-of-fit tests for the lognormal distribution: the Anderson-Darling, the Cramér–von Mises, and the Kolmogorov-Smirnov tests. The null hypothesis for the three tests is that a lognormal distribution holds for the sample data.
Output 4.27.2  Summary of Fitted Lognormal Distribution

<table>
<thead>
<tr>
<th>Distribution of Position Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>The UNIVARIATE Procedure</td>
</tr>
<tr>
<td>Fitted Lognormal Distribution for Deviation (Position Deviation)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters for Lognormal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Threshold</td>
</tr>
<tr>
<td>Scale</td>
</tr>
<tr>
<td>Shape</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Std Dev</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goodness-of-Fit Tests for Lognormal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
</tr>
<tr>
<td>Anderson-Darling</td>
</tr>
</tbody>
</table>

The $p$-values for all three tests are greater than 0.5, so the null hypothesis is not rejected. The tests support the conclusion that the two-parameter lognormal distribution with scale parameter $\xi = -6.14$ and shape parameter $\delta = 0.88$ provides a good model for the distribution of position deviations. For further discussion of goodness-of-fit interpretation, see the section “Goodness-of-Fit Tests” on page 413.

A sample program for this example, unix16.sas, is available in the SAS Sample Library for Base SAS software.

Example 4.28: Creating a Normal Quantile Plot

This example illustrates how to create a normal quantile plot. An engineer is analyzing the distribution of distances between holes cut in steel sheets. The following statements save measurements of the distance between two holes cut into 50 steel sheets as values of the variable Distance in the data set Sheets:

```sas
data Sheets;
  input Distance @@;
  label Distance = 'Hole Distance (cm)';
datalines;
9.80 10.20 10.27 9.70 9.76
10.11 10.24 10.20 10.24 9.63
9.99 9.78 10.10 10.21 10.00
9.96 9.79 10.08 9.79 10.06
10.10 9.95 9.84 10.11 9.93
10.56 10.47 9.42 10.44 10.16
10.11 10.36 9.94 9.77 9.36
```

The engineer decides to check whether the distribution of distances is normal. The following statements create a Q-Q plot for Distance, shown in Output 4.28.1:

```sas
symbol v=plus;
title 'Normal Quantile-Quantile Plot for Hole Distance';
ods graphics off;
proc univariate data=Sheets noprint;
  qqplot Distance;
run;
```

The plot compares the ordered values of Distance with quantiles of the normal distribution. The linearity of the point pattern indicates that the measurements are normally distributed. Note that a normal Q-Q plot is created by default.

**Output 4.28.1 Normal Quantile-Quantile Plot for Distance**

A sample program for this example, `uniex17.sas`, is available in the SAS Sample Library for Base SAS software.
Example 4.29: Adding a Distribution Reference Line

This example, which is a continuation of Example 4.28, illustrates how to add a reference line to a normal Q-Q plot, which represents the normal distribution with mean $\mu_0$ and standard deviation $\sigma_0$. The following statements reproduce the Q-Q plot in Output 4.28.1 and add the reference line:

```plaintext
symbol v=plus;
title 'Normal Quantile-Quantile Plot for Hole Distance';
ods graphics off;
proc univariate data=Sheets noprint;
   qqplot Distance / normal(mu=est sigma=est color=red l=2)
                          square;
run;
```

The plot is displayed in Output 4.29.1.

Specifying MU=EST and SIGMA=EST with the NORMAL primary option requests the reference line for which $\mu_0$ and $\sigma_0$ are estimated by the sample mean and standard deviation. Alternatively, you can specify numeric values for $\mu_0$ and $\sigma_0$ with the MU= and SIGMA= secondary options. The COLOR= and L= options specify the color and type of the line, and the SQUARE option displays the plot in a square format. The NOPRINT options in the PROC UNIVARIATE statement and after the NORMAL option suppress all the tables of statistical output produced by default.
The data clearly follow the line, which indicates that the distribution of the distances is normal.

A sample program for this example, `uniex17.sas`, is available in the SAS Sample Library for Base SAS software.
Example 4.30: Interpreting a Normal Quantile Plot

This example illustrates how to interpret a normal quantile plot when the data are from a non-normal distribution. The following statements create the data set Measures, which contains the measurements of the diameters of 50 steel rods in the variable Diameter:

```sas
data Measures;
   input Diameter @@;
   label Diameter = 'Diameter (mm)';
   datalines;
   5.501  5.251  5.404  5.366  5.445  5.576  5.607
   5.200  5.977  5.177  5.332  5.399  5.661  5.512
   5.252  5.404  5.739  5.525  5.160  5.410  5.823
   5.376  5.202  5.470  5.410  5.394  5.146  5.244
   5.309  5.480  5.388  5.399  5.360  5.368  5.394
   5.248  5.409  5.304  6.239  5.781  5.247  5.907
   5.208  5.143  5.304  5.603  5.164  5.209  5.475
   5.223
;```

The following statements request the normal Q-Q plot in Output 4.30.1:

```sas
symbol v=plus;
title 'Normal Q-Q Plot for Diameters';
ods graphics off;
proc univariate data=Measures noprint;
   qqplot Diameter / normal
      square
      vaxis=axis1;
      axis1 label=(a=90 r=0);
run;
```

The nonlinearity of the points in Output 4.30.1 indicates a departure from normality. Because the point pattern is curved with slope increasing from left to right, a theoretical distribution that is skewed to the right, such as a lognormal distribution, should provide a better fit than the normal distribution. The mild curvature suggests that you should examine the data with a series of lognormal Q-Q plots for small values of the shape parameter \( \sigma \), as illustrated in Example 4.31. For details on interpreting a Q-Q plot, see the section “Interpretation of Quantile-Quantile and Probability Plots” on page 420.
Chapter 4: The UNIVARIATE Procedure

Output 4.30.1 Normal Quantile-Quantile Plot of Nonnormal Data

A sample program for this example, uniex18.sas, is available in the SAS Sample Library for Base SAS software.

Example 4.31: Estimating Three Parameters from Lognormal Quantile Plots

This example, which is a continuation of Example 4.30, demonstrates techniques for estimating the shape, location, and scale parameters, and the theoretical percentiles for a three-parameter lognormal distribution.

The three-parameter lognormal distribution depends on a threshold parameter $\theta$, a scale parameter $\zeta$, and a shape parameter $\sigma$. You can estimate $\sigma$ from a series of lognormal Q-Q plots which use the SIGMA= secondary option to specify different values of $\sigma$; the estimate of $\sigma$ is the value that linearizes the point pattern. You can then estimate the threshold and scale parameters from the intercept and slope of the point pattern. The following statements create the series of plots in Output 4.31.1, Output 4.31.2, and Output 4.31.3 for $\sigma$ values of 0.2, 0.5, and 0.8, respectively:
symbol v=plus;
title 'Lognormal Q-Q Plot for Diameters';
ods graphics off;
proc univariate data=Measures noprint;
    qqplot Diameter / lognormal(sigma=0.2 0.5 0.8)
        square;
run;

NOTE: You must specify a value for the shape parameter $\sigma$ for a lognormal Q-Q plot with the SIGMA= option or its alias, the SHAPE= option.

Output 4.31.1 Lognormal Quantile-Quantile Plot ($\sigma = 0.2$)
Output 4.31.2 Lognormal Quantile-Quantile Plot ($\sigma = 0.5$)

Lognormal Q-Q Plot for Diameters
Example 4.31: Estimating Three Parameters from Lognormal Quantile Plots

**Output 4.31.3** Lognormal Quantile-Quantile Plot ($\sigma = 0.8$)

The plot in **Output 4.31.2** displays the most linear point pattern, indicating that the lognormal distribution with $\sigma = 0.5$ provides a reasonable fit for the data distribution.

Data with this particular lognormal distribution have the following density function:

$$p(x) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi} (x - \theta)} \exp \left(-2(\log(x - \theta) - \zeta)\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

The points in the plot fall on or near the line with intercept $\theta$ and slope $\exp(\xi)$. Based on **Output 4.31.2**, $\theta \approx 5$ and $\exp(\xi) \approx \frac{12}{\pi} \approx 0.4$, giving $\zeta \approx \log(0.4) \approx -0.92$.

You can also request a reference line by using the SIGMA=, THETA=, and ZETA= options together. The following statements produce the lognormal Q-Q plot in **Output 4.31.4**:

```plaintext
symbol v=plus;
title 'Lognormal Q-Q Plot for Diameters';
ods graphics off;
proc univariate data=Measures noprint;
    qqplot Diameter / lognormal(theta=5 zeta=est sigma=est
color=black l=2)
```
Output 4.31.1 through Output 4.31.3 show that the threshold parameter $\theta$ is not equal to zero. Specifying THETA=5 overrides the default value of zero. The SIGMA=EST and ZETA=EST secondary options request estimates for $\sigma$ and $\exp \zeta$ that use the sample mean and standard deviation.

**Output 4.31.4** Lognormal Quantile-Quantile Plot ($\sigma=$est, $\zeta=$est, $\theta=$5)

From the plot in Output 4.31.2, $\sigma$ can be estimated as 0.51, which is consistent with the estimate of 0.5 derived from the plot in Output 4.31.2. Example 4.32 illustrates how to estimate percentiles by using lognormal Q-Q plots.

A sample program for this example, uniex18.sas, is available in the SAS Sample Library for Base SAS software.

---

**Example 4.32: Estimating Percentiles from Lognormal Quantile Plots**

This example, which is a continuation of Example 4.31, shows how to use a Q-Q plot to estimate percentiles such as the 95th percentile of the lognormal distribution. A probability plot can also be used for this purpose, as illustrated in Example 4.26.
Example 4.32: Estimating Percentiles from Lognormal Quantile Plots

The point pattern in Output 4.31.4 has a slope of approximately 0.39 and an intercept of 5. The following statements reproduce this plot, adding a lognormal reference line with this slope and intercept:

```plaintext
symbol v=plus;
title 'Lognormal Q-Q Plot for Diameters';
ods graphics off;
proc univariate data=Measures noprint;
   qqplot Diameter / lognormal(sigma=0.5 theta=5 slope=0.39)
      pctlaxis(grid)
      vref = 5.8 5.9 6.0
      square;
run;
```

The result is shown in Output 4.32.1:

**Output 4.32.1** Lognormal Q-Q Plot Identifying Percentiles

The PCTLAXIS option labels the major percentiles, and the GRID option draws percentile axis reference lines. The 95th percentile is 5.9, because the intersection of the distribution reference line and the 95th reference line occurs at this value on the vertical axis.

Alternatively, you can compute this percentile from the estimated lognormal parameters. The $\alpha$th percentile
of the lognormal distribution is

\[ P_\alpha = \exp(\sigma \Phi^{-1}(\alpha) + \xi) + \theta \]

where \( \Phi^{-1}(\cdot) \) is the inverse cumulative standard normal distribution. Consequently,

\[ \hat{P}_{0.95} = \exp\left( \frac{1}{2} \Phi^{-1}(0.95) + \log(0.39) \right) + 5 = 5.89 \]

A sample program for this example, `uniex18.sas`, is available in the SAS Sample Library for Base SAS software.

---

**Example 4.33: Estimating Parameters from Lognormal Quantile Plots**

This example, which is a continuation of Example 4.31, demonstrates techniques for estimating the shape, location, and scale parameters, and the theoretical percentiles for a two-parameter lognormal distribution.

If the threshold parameter is known, you can construct a two-parameter lognormal Q-Q plot by subtracting the threshold from the data values and making a normal Q-Q plot of the log-transformed differences, as illustrated in the following statements:

```sas
data ModifiedMeasures;
  set Measures;
  LogDiameter = log(Diameter-5);
  label LogDiameter = 'log(Diameter-5)';
run;

symbol v=plus;
title 'Two-Parameter Lognormal Q-Q Plot for Diameters';
ods graphics off;
proc univariate data=ModifiedMeasures noprint;
  qqplot LogDiameter / normal(mu=est sigma=est)
    square
    vaxis=axis1;
  inset n mean (5.3) std (5.3)
    / pos = nw header = 'Summary Statistics';
    axis1 label=(a=90 x=0);
run;
```
Example 4.33: Estimating Parameters from Lognormal Quantile Plots

Output 4.33.1 Two-Parameter Lognormal Q-Q Plot for Diameters

Because the point pattern in Output 4.33.1 is linear, you can estimate the lognormal parameters \( \xi \) and \( \sigma \) as the normal plot estimates of \( \mu \) and \( \sigma \), which are \(-0.99\) and \(0.51\). These values correspond to the previous estimates of \(-0.92\) for \( \xi \) and \(0.5\) for \( \sigma \) from Example 4.31. A sample program for this example, `uniex18.sas`, is available in the SAS Sample Library for Base SAS software.
Example 4.34: Comparing Weibull Quantile Plots

This example compares the use of three-parameter and two-parameter Weibull Q-Q plots for the failure times in months for 48 integrated circuits. The times are assumed to follow a Weibull distribution. The following statements save the failure times as the values of the variable Time in the data set Failures:

```sas
data Failures;
    input Time @@;
    label Time = 'Time in Months';
    datalines;
29.42 32.14 30.58 27.50 26.08 29.06 25.10 31.34
29.14 33.96 30.64 27.32 29.86 26.28 29.68 33.76
29.32 30.82 27.26 27.92 30.92 24.64 32.90 35.46
26.72 27.42 29.02 27.54 31.60 33.46 26.78 27.82
29.18 27.94 27.66 26.42 31.00 26.64 31.44 32.52
;
```

If no assumption is made about the parameters of this distribution, you can use the WEIBULL option to request a three-parameter Weibull plot. As in the previous example, you can visually estimate the shape parameter $c$ by requesting plots for different values of $c$ and choosing the value of $c$ that linearizes the point pattern. Alternatively, you can request a maximum likelihood estimate for $c$, as illustrated in the following statements:

```sas
symbol v=plus;
title 'Three-Parameter Weibull Q-Q Plot for Failure Times';
ods graphics off;
proc univariate data=Failures noprint;
    qqplot Time / weibull(c=est theta=est sigma=est) square
        href=0.5 1 1.5 2
        vref=25 27.5 30 32.5 35
        lhref=4 lvref=4;
run;
```

**NOTE:** When using the WEIBULL option, you must either specify a list of values for the Weibull shape parameter $c$ with the C= option or specify C=EST.

Output 4.34.1 displays the plot for the estimated value $\hat{c} = 1.99$. The reference line corresponds to the estimated values for the threshold and scale parameters of $\hat{\theta} = 24.19$ and $\hat{\sigma} = 5.83$, respectively.
Now, suppose it is known that the circuit lifetime is at least 24 months. The following statements use the known threshold value $\theta_0 = 24$ to produce the two-parameter Weibull Q-Q plot shown in Output 4.31.4:

```plaintext
symbol v=plus;
title 'Two-Parameter Weibull Q-Q Plot for Failure Times';
ods graphics off;
proc univariate data=Failures noprint;
  qqplot Time / weibull(theta=24 c=est sigma=est)
square
  vref= 25 to 35 by 2.5
  href= 0.5 to 2.0 by 0.5
  lhref=4 lvref=4;
run;
```

The reference line is based on maximum likelihood estimates $\hat{\theta} = 2.08$ and $\hat{\sigma} = 6.05$. 
Example 4.35: Creating a Cumulative Distribution Plot

A company that produces fiber-optic cord is interested in the breaking strength of the cord. The following statements create a data set named Cord, which contains 50 breaking strengths measured in pounds per square inch (PSI):

```sas
data Cord;
  label Strength="Breaking Strength (psi)";
  input Strength @@;
datalines;
6.94 6.97 7.11 6.95 7.12 6.70 7.13 7.34 6.90 6.83
7.06 6.89 7.28 6.93 7.05 7.00 7.04 7.21 7.08 7.01
7.05 7.11 7.03 6.98 7.04 7.08 6.87 6.81 7.11 6.74
6.95 7.05 6.98 6.94 7.06 7.12 7.19 7.12 7.01 6.84
6.91 6.89 7.23 6.98 6.93 6.83 6.99 7.00 6.97 7.01
;```

A sample program for this example, `uniex19.sas`, is available in the SAS Sample Library for Base SAS software.
You can use the CDFPLOT statement to fit any of six theoretical distributions (beta, exponential, gamma, lognormal, normal, and Weibull) and superimpose them on the cdf plot. The following statements use the NORMAL option to display a fitted normal distribution function on a cdf plot of breaking strengths:

```plaintext
title 'Cumulative Distribution Function of Breaking Strength';
ods graphics on;
proc univariate data=Cord noprint;
   cdf Strength / normal;
   inset normal(mu sigma);
run;
```

The NORMAL option requests the fitted curve. The INSET statement requests an inset containing the parameters of the fitted curve, which are the sample mean and standard deviation. For more information about the INSET statement, see “INSET Statement” on page 316. The resulting plot is shown in Output 4.35.1.

**Output 4.35.1** Cumulative Distribution Function

The plot shows a symmetric distribution with observations concentrated at 6.9 and 7.1. The agreement between the empirical and the normal distribution functions in Output 4.35.1 is evidence that the normal distribution is an appropriate model for the distribution of breaking strengths.
Example 4.36: Creating a P-P Plot

The distances between two holes cut into 50 steel sheets are measured and saved as values of the variable Distance in the following data set:

```sas
data Sheets;
    input Distance @@;
    label Distance='Hole Distance in cm';
datalines;
  9.80 10.20 10.27 9.70 9.76
  10.11 10.24 10.20 10.24 9.63
  9.99 9.78 10.10 10.21 10.00
  9.96 9.79 10.08 9.79 10.06
  10.10 9.95 9.84 10.11 9.93
  10.56 10.47 9.42 10.44 10.16
  10.11 10.36 9.94 9.77 9.36
  9.89 9.62 10.05 9.72 9.82
  9.99 10.16 10.58 10.70 9.54
  10.31 10.07 10.33 9.98 10.15
;
```

It is decided to check whether the distances are normally distributed. The following statements create a P-P plot, shown in Output 4.36.1, which is based on the normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 0.3$:

```sas
title 'Normal Probability-Probability Plot for Hole Distance';
ods graphics on;
proc univariate data=Sheets noprint;
    ppplot Distance / normal(mu=10 sigma=0.3)
        square;
run;
```

The NORMAL option in the PPPLOT statement requests a P-P plot based on the normal cumulative distribution function, and the MU= and SIGMA= normal-options specify $\mu$ and $\sigma$. Note that a P-P plot is always based on a completely specified distribution—in other words, a distribution with specific parameters. In this example, if you did not specify the MU= and SIGMA= normal-options, the sample mean and sample standard deviation would be used for $\mu$ and $\sigma$. 
The linearity of the pattern in Output 4.36.1 is evidence that the measurements are normally distributed with mean 10 and standard deviation 0.3. The SQUARE option displays the plot in a square format.
References


Subject Index

adjusted odds ratio
  FREQ procedure, 201
adjusted relative risks
  FREQ procedure, 202
agreement plots
  FREQ procedure, 82
agreement, measures of
  FREQ procedure, 192
Agresti-Caffo confidence limits
  risk difference (FREQ), 176
Agresti-Coull confidence limits
  proportions (FREQ), 168
Anderson-Darling statistic, 416
Anderson-Darling test, 278
annotating
  histograms, 369
ANOVA (row mean scores) statistic
  Mantel-Haenszel (FREQ), 199
association, measures of
  FREQ procedure, 158
bar charts
  FREQ procedure, 135
Barnard’s test
  FREQ procedure, 184
beta distribution, 401, 422
cdf plots, 288
deviation from theoretical distribution, 415
EDF goodness-of-fit test, 414
estimation of parameters, 305
fitting, 305, 401
formulas for, 401
P-P plots, 337
probability plots, 350, 422
quantile plots, 361, 422
binomial proportions
  Clopper-Pearson test (FREQ), 171
close confidence limits (FREQ), 168
equivalence tests (FREQ), 173
exact test (FREQ), 171
FREQ procedure, 168
noninferiority tests (FREQ), 171
superiority tests (FREQ), 172
tests (FREQ), 170
TOST (FREQ), 173
Bowker’s test of symmetry
  FREQ procedure, 192, 193
box plots, line printer, 278, 390
  side-by-side, 278, 391
Breslow-Day test
  FREQ procedure, 202
  Tarone’s adjustment (FREQ), 202
case-control studies
  odds ratio (FREQ), 186
categorical data analysis
  FREQ procedure, 70
cdf plots, 282
  axes, specifying, 293
  beta distribution, 288
  creating, 518
  example, 518
  exponential distribution, 289
  gamma distribution, 289
  gumbel distribution, 290
  igauss distribution, 290
  lognormal distribution, 291
  normal distribution, 291
  normal distribution, example, 518
  pareto distribution, 292
  power function distribution, 292
  rayleigh distribution, 293
  suppressing empirical cdf, 291
  Weibull distribution, 294
cell count data
  example (FREQ), 227
  FREQ procedure, 146
chi-square goodness-of-fit test
  FREQ procedure, 153
chi-square tests
  FREQ procedure, 153
Cicchetti-Allison weights
  kappa coefficient (FREQ), 195
Clopper-Pearson confidence limits
  proportions (FREQ), 169
Cochran’s $Q$ test
  FREQ procedure, 192, 196
Cochran-Armitage test for trend
  FREQ procedure, 189
Cochran-Mantel-Haenszel statistics
  FREQ procedure, 197
cohort studies
  relative risks (FREQ), 188
common odds ratio
  exact confidence limits (FREQ), 203
  exact test (FREQ), 203
logit (FREQ), 201
Mantel-Haenszel (FREQ), 201
common relative risks
logit (FREQ), 202
Mantel-Haenszel (FREQ), 202
comparative plots, 294–296, 396
histograms, 310, 311, 373, 374, 466, 469, 479
concordant observations
FREQ procedure, 158
confidence ellipse, 32
confidence limits
exact (FREQ), 85
for percentiles, 383
means, for, 386
measures of association (FREQ), 159
parameters of normal distribution, for, 386
proportions (FREQ), 168
standard deviations, for, 386
variances, for, 386
confidence limits for the correlation
Fisher’s z transformation, 26
contingency coefficient
FREQ procedure, 158
contingency tables
FREQ procedure, 70, 104
continuity-adjusted chi-square test
FREQ procedure, 156
CORR procedure
concepts, 19
details, 19
examples, 38
missing values, 33
ODS graph names, 38
ODS table names, 37
output, 34
output data sets, 35
overview, 4
results, 33
syntax, 9
task tables, 9
corrected sums of squares and crossproducts, 10
correlation coefficients, 4
limited combinations of, 19
printing, for each variable, 10
suppressing probabilities, 10
correlation statistic
Mantel-Haenszel (FREQ), 198
covariances, 10
Cramér’s V statistic
FREQ procedure, 158
Cramér–von Mises statistic, 416
Cramér–von Mises test, 278
Cronbach’s coefficient alpha, 31
calculating and printing, 10
example, 56
for estimating reliability, 4
crosstabulation tables
FREQ procedure, 70, 104, 216
data summarization tools, 266
density estimation, see kernel density estimation
descriptive statistics
computing, 379
discordant observations
FREQ procedure, 158
distribution of variables, 266
dot plots
FREQ procedure, 135, 230
EDF, see empirical distribution function
EDF goodness-of-fit tests, 414
probability values of, 416
empirical distribution function
definition of, 415
EDF test statistics, 415
equivalence tests
binomial proportions (FREQ), 173
risk difference (FREQ), 182
exact confidence limits
odds ratio (FREQ), 187
proportion difference (FREQ), 183
proportions (FREQ), 169
ratio of proportions (FREQ), 188
relative risks (FREQ), 188
risk difference (FREQ), 183
exact p-values
FREQ procedure, 207
exact tests
computational algorithms (FREQ), 207
computational resources (FREQ), 208
FREQ procedure, 85, 206, 248
Monte Carlo estimation (FREQ), 92
network algorithm (FREQ), 207
exponential distribution, 403, 423
cdf plots, 289
development from theoretical distribution, 415
EDF goodness-of-fit test, 414
estimation of parameters, 307
fitting, 403
formulas for, 403
P-P plots, 338
probability plots, 351, 423
quantile plots, 362, 423
extreme observations, 316, 443
extreme values, 443
Farrington-Manning test
risk difference (FREQ), 181
Fisher’s exact test
FREQ procedure, 157
Fisher’s z transformation
  applications, 27
Fisher’s z transformation, 10, 25
  confidence limits for the correlation, 26
fitted parametric distributions, 401
  beta distribution, 401
  exponential distribution, 403
  folded normal distribution, 491
  gamma distribution, 403
  gumbel distribution, 404
  inverse Gaussian distribution, 405
  Johnson S distribution, 410
  Johnson SU distribution, 411
  lognormal distribution, 406
  normal distribution, 406
  pareto distribution, 407
  power function distribution, 408
  rayleigh distribution, 409
  Weibull distribution, 413
Fleiss-Cohen weights
  kappa coefficient (FREQ), 195
folded normal distribution, 491
Freeman-Halton test
  FREQ procedure, 157
FREQ procedure
  adjusted odds ratio (Mantel-Haenszel), 201
  adjusted relative risks (Mantel-Haenszel), 202
  Agresti-Caffo confidence limits, 176
  Agresti-Coull confidence limits, 168
  ANOVA (row mean scores) statistic, 199
  bar charts, 135
  Barnard’s test, 184
  binomial proportions, 168
  Bowker’s test of symmetry, 192, 193
  Breslow-Day test, 202
  cell count data, 146
  chi-square goodness-of-fit test, 153
  chi-square tests, 153
  Clopper-Pearson confidence limits, 169
  Cochran’s $Q$ test, 192, 196
  Cochran-Armitage test for trend, 189
  common odds ratio, 203
  computational resources, 210
  computational resources (exact tests), 208
  contingency coefficient, 158
  continuity-adjusted chi-square test, 156
  correlation statistic, 198
  Cramér’s $V$ statistic, 158
  crosstabulation tables, 216
  default tables, 104
  displayed output, 214
  dot plots, 135, 230
  equivalence tests, 173
  equivalence tests (risk difference), 182
  exact confidence limits, 85
  exact $p$-values, 207
  exact tests, 85, 206, 248
  exact unconditional confidence limits, 183
  Farrington-Manning test, 181
  Fisher’s exact test, 157
  Freeman-Halton test, 157
  Friedman’s chi-square test, 251
  Gail-Simon test, 205
  gamma statistic, 158, 160
  general association statistic, 199
  grouping with formats, 147
  Hauck-Anderson confidence limits, 177
  in-database computation, 150
  input data sets, 83, 146
  introductory examples, 72
  Jeffreys confidence limits, 169
  Jonckheere-Terpstra test, 190
  kappa coefficient, 192, 193
  Kendall’s tau-b statistic, 158, 160
  lambda asymmetric, 158, 165
  lambda symmetric, 158, 166
  likelihood ratio chi-square test, 156
  Mantel-Fleiss criterion, 200
  Mantel-Haenszel chi-square test, 156
  Mantel-Haenszel statistics, 197
  maximum time (exact tests), 91
  McNemar’s test, 192
  measures of agreement, 192
  measures of association, 158
  Miettinen-Nurminen confidence limits, 177
  missing values, 148
  Monte Carlo estimation (exact tests), 85, 92, 209
  mosaic plots, 128
  multiway tables, 216
  network algorithm, 207
  Newcombe confidence limits, 178, 181
  noninferiority tests, 171
  noninferiority tests (risk difference), 180
  odds ratio, 186
  ODS graph names, 226
  ODS table names, 222
  one-way frequency tables, 214
  ordering of levels, 84
  output data sets, 93, 211
  overall kappa coefficient, 196
  Pearson chi-square test, 154
  Pearson correlation coefficient, 158, 162
  phi coefficient, 158
  polychoric correlation coefficient, 158, 164
  relative risks, 188
  risk difference, 174
  score confidence limits, 177, 187
scores, 152
simple kappa coefficient, 193
Somers’ D statistics, 158, 161
Spearman rank correlation coefficient, 158, 163
standardized residuals, 154
Stuart’s tau-c statistic, 158, 161
superiority tests, 172
superiority tests (risk difference), 182
tetrachoric correlation coefficient, 164
uncertainty coefficients, 158, 167
Wald confidence limits (risk difference), 179
weighted kappa coefficient, 192, 194
Wilson confidence limits, 169
Yule’s Q statistic, 160
Zelen’s exact test, 203

frequency plots
 FREQ procedure, 74

frequency tables
 creating (UNIVARIATE), 445
 FREQ procedure, 70, 104
 one-way (FREQ), 214

Friedman’s chi-square test
 FREQ procedure, 251

Gail-Simon test
 FREQ procedure, 205

gamma distribution, 403, 423
cdf plots, 289
development from theoretical distribution, 415
EDF goodness-of-fit test, 414
estimation of parameters, 308
fitting, 308, 403
formulas for, 403
P-P plots, 338, 340
probability plots, 351, 423
quantile plots, 362, 423

gamma statistic
 FREQ procedure, 158, 160
general association statistic
 Mantel-Haenszel (FREQ), 199

Gini’s mean difference, 388
goodness-of-fit tests, 278, 413, see empirical distribution function, 483
 Anderson-Darling, 416
 Cramér–von Mises, 416
 EDF, 414, 416
 Kolmogorov D, 415
 Shapiro-Wilk, 414

graphics, 266
 annotating, 276
 descriptions, 371
 high-resolution, 266, 391
 insets, 316, 397–399
 line printer, 389

naming, 373
 probability plots, 345
 quantile plots, 356
 saving, 277

Gumbel distribution, 424
cdf plots, 290
estimation of parameters, 308
fitting, 308, 404
formulas for, 404
probability plots, 352
quantile plots, 363

Hauck-Anderson confidence limits
 risk difference (FREQ), 177

high-resolution graphics, 266, 391

histograms, 296, 437
adding a grid, 308
annotating, 369
appearance, 304–306, 309, 310, 312, 314, 315, 369–374, 376, 377
axis color, 369
axis scaling, 315
bar labels, 304
bar width, 304, 311
bars, suppressing, 312
beta curve, superimposed, 305
binning, 473
color, options, 305, 306, 369–371
comparative, 310, 311, 373, 374, 466, 469, 479
creating, 464
endpoints of intervals, 314
exponential curve, superimposed, 307
extreme observations, 443
filling area under density curve, 308
gamma curve, superimposed, 308
Gumbel curve, superimposed, 308
hanging, 308
insets, 471
intervals, 313, 437
inverse Gaussian curve, superimposed, 309
Johnson SB curve, superimposed, 314
Johnson SU curve, superimposed, 315
kernel density estimation, options, 305, 310, 315
kernel density estimation, superimposed, 418, 488
line type, 373
lognormal curve, superimposed, 310
lognormal distribution, 501
midpoints, 311
multiple distributions, example, 483
normal curve, superimposed, 312
Subject Index

normal distribution, 312
output data sets, 431, 432
parameters for fitted density curves, 304, 305, 312, 314–316
pareto curve, superimposed, 313
plots, suppressing, 312
power curve, superimposed, 313
quantiles, 437
rayleigh curve, superimposed, 314
reference lines, options, 308, 370–373, 377
saving histogram intervals, 313
tables of statistical output, 437
tables of statistical output, suppressing, 312
three-parameter lognormal distribution, superimposed, 489
three-parameter Weibull distribution, superimposed, 490
tick marks on horizontal axis, 371
tiles for comparative plots, 372
Weibull curve, superimposed, 315
Hoeffding’s measure of dependence, 4, 23
calculating and printing, 10
example, 38
output data set with, 10
probability values, 23
hypothesis tests
exact (FREQ), 85
igauss distribution
cdf plots, 290
in-database computation
FREQ procedure, 150
insets, 316, 471
appearance, 325
appearance, color, 324
positioning, 325, 397
positioning in margins, 398
positioning with compass point, 397
positioning with coordinates, 399
statistics associated with distributions, 320
insets for descriptive statistics, see insets
interquartile range, 388
inverse Gaussian distribution, 405
estimation of parameters, 309
fitting, 309, 405
formulas for, 405
P-P plots, 340
Jeffreys confidence limits
proportions (FREQ), 169
Johnson $S_B$ distribution, 410
estimation of parameters, 314
fitting, 314, 410
formulas for, 410
Johnson $S_U$ distribution, 411
estimation of parameters, 315
fitting, 315, 411
formulas for, 411
Jonckheere-Terpstra test
FREQ procedure, 190
kappa coefficient
FREQ procedure, 192, 193
weights (FREQ), 195
Kendall correlation statistics, 10
Kendall’s partial tau-b, 4, 18
Kendall’s tau-b, 4, 22
probability values, 22
Kendall’s tau-b statistic
FREQ procedure, 158, 160
kernel density estimation, 418, 488
adding density curve to histogram, 310
bandwidth parameter, specifying, 305
kernel function, specifying type of, 310
line type for density curve, 373
lower bound, specifying, 310
upper bound, specifying, 315
kernel function, see kernel density estimation
key cell for comparative plots, 296
Kolmogorov D statistic, 415
Kolmogorov-Smirnov test, 278
lambda asymmetric
FREQ procedure, 158, 165
lambda symmetric
FREQ procedure, 158, 166
likelihood ratio chi-square test
FREQ procedure, 156
line printer plots, 389
box plots, 390, 391
normal probability plots, 390
stem-and-leaf plots, 390
listwise deletion, 33
location estimates
robust, 281
location parameters, 427
probability plots, estimation with, 427
quantile plots, estimation with, 427
location, tests for
UNIVARIATE procedure, 462
lognormal distribution, 406, 424
cdf plots, 291
deviation from theoretical distribution, 415
EDF goodness-of-fit test, 414
 estimation of parameters, 310
fitting, 310, 406
formulas for, 406
histograms, 489, 501
P-P plots, 341
probability plots, 352, 424, 497
quantile plots, 363, 424, 514

Mantel-Fleiss criterion
FREQ procedure, 200
Mantel-Haenszel chi-square test
FREQ procedure, 156
Mantel-Haenszel statistics
ANOVA (row mean scores) statistic (FREQ), 199
correlation statistic (FREQ), 198
FREQ procedure, 197
general association statistic (FREQ), 199
Mantel-Fleiss criterion (FREQ), 200

McNemar’s test
FREQ procedure, 192
measures of association, 38
exact tests (FREQ), 160
nonparametric, 4
tests (FREQ), 159
measures of location
means, 379
modes, 382, 441
trimmed means, 388
Winsorized means, 387
median absolute deviation (MAD), 388
Mehta-Patel network algorithm
exact tests (FREQ), 207
Miettinen-Nurminen confidence limits
risk difference (FREQ), 177
missing values
FREQ procedure, 148
UNIVARIATE procedure, 377
mode calculation, 382
modified ridit scores
FREQ procedure, 152
Monte Carlo estimation
exact tests (FREQ), 85, 92, 209
mosaic plots
FREQ procedure, 128
multiway tables
FREQ procedure, 70, 104, 216
network algorithm
exact tests (FREQ), 207
Newcombe confidence limits
risk difference (FREQ), 178, 181
Newton-Raphson approximation
gamma shape parameter, 369
noninferiority tests
binomial proportions (FREQ), 171
risk difference (FREQ), 180
nonparametric density estimation, see kernel density estimation

nonparametric measures of association, 4
normal distribution, 406, 424
cdf plots, 291
cdf plots, example, 518
deviation from theoretical distribution, 415
EDF goodness-of-fit test, 414
estimation of parameters, 312
fitting, 312, 406
formulas for, 406
histograms, 312
P-P plots, 342
probability plots, 346, 352, 424
quantile plots, 363, 424, 503
normal probability plots, see probability plots
line printer, 278, 390

odds ratio
Breslow-Day test (FREQ), 202
case-control studies (FREQ), 186
exact confidence limits (FREQ), 187
FREQ procedure, 186
logit adjusted (FREQ), 201
Mantel-Haenszel adjusted (FREQ), 201
Zelen’s exact test (FREQ), 203

ODS (Output Delivery System)
CORR procedure and, 37
UNIVARIATE procedure table names, 436

ODS graph names
CORR procedure, 38
FREQ procedure, 226

output data sets
saving correlations in, 59

P-P plots, 332
beta distribution, 337
distribution options, 333, 335
distribution reference line, 334, 521
exponential distribution, 338
gamma distribution, 338, 339
Gumbel distribution, 339
gumbel distribution, 339
inverse Gaussian distribution, 340
lognormal distribution, 341
normal distribution, 342
options summarized by function, 335, 336
Pareto distribution, 342
plot layout, 335
Power distribution, 343
Rayleigh distribution, 343
Weibull distribution, 344

paired data, 384, 463
pairwise deletion, 33
parameters for fitted density curves, 304, 305, 312, 314–316
<table>
<thead>
<tr>
<th>Subject Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto distribution, 425</td>
</tr>
<tr>
<td>P-P plots, 342</td>
</tr>
<tr>
<td>probability plots, 425</td>
</tr>
<tr>
<td>quantile plots, 425</td>
</tr>
<tr>
<td>pareto distribution, 407</td>
</tr>
<tr>
<td>cdf plots, 292</td>
</tr>
<tr>
<td>estimation of parameters, 313</td>
</tr>
<tr>
<td>fitting, 313, 407</td>
</tr>
<tr>
<td>formulas for, 407</td>
</tr>
<tr>
<td>probability plots, 353</td>
</tr>
<tr>
<td>quantile plots, 363</td>
</tr>
<tr>
<td>partial correlations, 23</td>
</tr>
<tr>
<td>probability values, 25</td>
</tr>
<tr>
<td>Pearson chi-square test</td>
</tr>
<tr>
<td>FREQ procedure, 154</td>
</tr>
<tr>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>FREQ procedure, 158, 162</td>
</tr>
<tr>
<td>Pearson correlation statistics, 4</td>
</tr>
<tr>
<td>example, 38</td>
</tr>
<tr>
<td>in output data set, 10</td>
</tr>
<tr>
<td>Pearson partial correlation, 4, 18</td>
</tr>
<tr>
<td>Pearson product-moment correlation, 4, 10, 19, 38</td>
</tr>
<tr>
<td>Pearson weighted product-moment correlation, 4, 18</td>
</tr>
<tr>
<td>probability values, 21</td>
</tr>
<tr>
<td>suppressing, 10</td>
</tr>
<tr>
<td>percent plots, see P-P plots</td>
</tr>
<tr>
<td>percentiles</td>
</tr>
<tr>
<td>axes, quantile plots, 364, 428</td>
</tr>
<tr>
<td>calculating, 382</td>
</tr>
<tr>
<td>confidence limits for, 383, 457</td>
</tr>
<tr>
<td>defining, 278, 382</td>
</tr>
<tr>
<td>empirical distribution function, 382</td>
</tr>
<tr>
<td>options, 330, 331</td>
</tr>
<tr>
<td>probability plots and, 345</td>
</tr>
<tr>
<td>quantile plots and, 356</td>
</tr>
<tr>
<td>saving to an output data set, 454</td>
</tr>
<tr>
<td>visual estimates, probability plots, 428</td>
</tr>
<tr>
<td>visual estimates, quantile plots, 428</td>
</tr>
<tr>
<td>weighted, 383</td>
</tr>
<tr>
<td>weighted average, 382</td>
</tr>
<tr>
<td>phi coefficient</td>
</tr>
<tr>
<td>FREQ procedure, 158</td>
</tr>
<tr>
<td>plot statements, UNIVARIATE procedure, 265</td>
</tr>
<tr>
<td>plots</td>
</tr>
<tr>
<td>box plots, 278, 390, 391</td>
</tr>
<tr>
<td>comparative, 294–296, 396</td>
</tr>
<tr>
<td>comparative histograms, 310, 311, 373, 374, 466, 469, 479</td>
</tr>
<tr>
<td>line printer plots, 389</td>
</tr>
<tr>
<td>normal probability plots, 278, 390</td>
</tr>
<tr>
<td>probability plots, 345, 421</td>
</tr>
<tr>
<td>quantile plots, 356, 421</td>
</tr>
<tr>
<td>reference lines, options, 376</td>
</tr>
<tr>
<td>size of, 280</td>
</tr>
<tr>
<td>stem-and-leaf, 278, 390</td>
</tr>
<tr>
<td>suppressing, 312</td>
</tr>
<tr>
<td>polychoric correlation, 28</td>
</tr>
<tr>
<td>polychoric correlation coefficient</td>
</tr>
<tr>
<td>FREQ procedure, 158, 164</td>
</tr>
<tr>
<td>polyserial correlation, 29</td>
</tr>
<tr>
<td>polyserial correlation statistics, 4, 10</td>
</tr>
<tr>
<td>Power distribution</td>
</tr>
<tr>
<td>P-P plots, 343</td>
</tr>
<tr>
<td>power function distribution, 408, 425</td>
</tr>
<tr>
<td>cdf plots, 292</td>
</tr>
<tr>
<td>estimation of parameters, 313</td>
</tr>
<tr>
<td>fitting, 313, 408</td>
</tr>
<tr>
<td>formulas for, 408</td>
</tr>
<tr>
<td>probability plots, 353, 425</td>
</tr>
<tr>
<td>quantile plots, 364, 425</td>
</tr>
<tr>
<td>prediction ellipse, 32</td>
</tr>
<tr>
<td>probability plots, 345</td>
</tr>
<tr>
<td>appearance, 351–353</td>
</tr>
<tr>
<td>beta distribution, 350, 422</td>
</tr>
<tr>
<td>distribution reference lines, 354</td>
</tr>
<tr>
<td>distributions for, 421</td>
</tr>
<tr>
<td>exponential distribution, 351, 423</td>
</tr>
<tr>
<td>gamma distribution, 351, 423</td>
</tr>
<tr>
<td>generalized Pareto distribution, 425</td>
</tr>
<tr>
<td>Gumbel distribution, 424</td>
</tr>
<tr>
<td>gumbel distribution, 352</td>
</tr>
<tr>
<td>location parameters, estimation of, 427</td>
</tr>
<tr>
<td>lognormal distribution, 352, 424, 497, 501</td>
</tr>
<tr>
<td>normal distribution, 346, 352, 424</td>
</tr>
<tr>
<td>overview, 345</td>
</tr>
<tr>
<td>parameters for distributions, 350–352, 354–356</td>
</tr>
<tr>
<td>Pareto distribution, 425</td>
</tr>
<tr>
<td>pareto distribution, 353</td>
</tr>
<tr>
<td>percentile axis, 353</td>
</tr>
<tr>
<td>percentiles, estimates of, 428</td>
</tr>
<tr>
<td>percentile axis, 353</td>
</tr>
<tr>
<td>scale parameters, estimation of, 427</td>
</tr>
<tr>
<td>shape parameters, 427</td>
</tr>
<tr>
<td>three-parameter Weibull distribution, 426</td>
</tr>
<tr>
<td>threshold parameter, 355</td>
</tr>
<tr>
<td>threshold parameters, estimation of, 427</td>
</tr>
<tr>
<td>two-parameter Weibull distribution, 426</td>
</tr>
<tr>
<td>Weibull distribution, 355</td>
</tr>
<tr>
<td>proportion difference</td>
</tr>
<tr>
<td>FREQ procedure, 174</td>
</tr>
<tr>
<td>proportions, see binomial proportions (FREQ)</td>
</tr>
<tr>
<td>Q-Q plots, see quantile plots</td>
</tr>
</tbody>
</table>
quantile plots, 356
appearance, 362, 363, 366
axes, percentile scale, 364, 428
beta distribution, 361, 422
creating, 420
diagnostics, 421
distribution reference lines, 365, 505
distributions for, 421
exponential distribution, 362, 423
gamma distribution, 362, 423
Gumbel distribution, 424
gumbel distribution, 363
interpreting, 420
legends, suppressing (UNIVARIATE), 505
location parameters, estimation of, 427
lognormal distribution, 363, 424, 508, 514
lognormal distribution, percentiles, 512
nonnormal data, 507
normal distribution, 363, 424, 503
overview, 356
parameters for distributions, 361–363, 365–367
Pareto distribution, 425
pareto distribution, 363
percentiles, estimates of, 428
power function distribution, 364, 425
Rayleigh distribution, 425
rayleigh distribution, 364
reference lines, 362, 364
reference lines, options, 363
scale parameters, estimation of, 427
shape parameters, 427
three-parameter Weibull distribution, 426
threshold parameter, 366
threshold parameters, estimation of, 427
two-parameter Weibull distribution, 426
Weibull distribution, 366, 516
quantile-quantile plots, see quantile plots
quantiles
defining, 382
empirical distribution function, 382
histograms and, 437
weighted average, 382
rank scores
FREQ procedure, 152
Rayleigh distribution, 425
P-P plots, 343
probability plots, 427
quantile plots, 427
rayleigh distribution, 409
cdf plots, 293
estimation of parameters, 314
fitting, 314, 409
formulas for, 409
probability plots, 353
quantile plots, 364
relative risks
cohort studies (FREQ), 188
exact confidence limits (FREQ), 188
FREQ procedure, 188
logit adjusted (FREQ), 202
Mantel-Haenszel adjusted (FREQ), 202
reliability estimation, 4
ridit scores
FREQ procedure, 152
risk difference
confidence limits (FREQ), 176
equivalence tests (FREQ), 182
exact confidence limits (FREQ), 183
FREQ procedure, 174
noninferiority tests (FREQ), 180
superiority tests (FREQ), 182
tests (FREQ), 179
TOST (FREQ), 182
risks, see also binomial proportions (FREQ)
FREQ procedure, 174
robust estimates, 280, 281
robust estimators, 387, 460
Gini’s mean difference, 388
interquartile range, 388
median absolute deviation (MAD), 388
$Q_n$, 388
$S_n$, 388
trimmed means, 388
Winsorized means, 387
robust measures of scale, 388
rounding, 280, 378
UNIVARIATE procedure, 378
row mean scores statistic
Mantel-Haenszel (FREQ), 199
saving correlations
example, 59
scale estimates
robust, 280
scale parameters, 427
probability plots, 427
quantile plots, 427
score confidence limits
odds ratio (FREQ), 187
risk difference (FREQ), 177
shape parameters, 427
Shapiro-Wilk statistic, 414
Shapiro-Wilk test, 278
sign test, 384, 385
paired data and, 463
signed rank statistic, computing, 385
singularity of variables, 10
smoothing data distribution, see kernel density estimation
$S_n$, 388
Somers’ $D$ statistics
  FREQ procedure, 158, 161
Spearman correlation statistics, 10
  probability values, 21
  Spearman partial correlation, 4, 18
  Spearman rank-order correlation, 4, 21, 38
Spearman rank correlation coefficient
  FREQ procedure, 158, 163
standard deviation, 10
  specifying, 291
standardized residuals
  FREQ procedure, 154
stem-and-leaf plots, 278, 390
stratified analysis
  FREQ procedure, 70, 104
Stuart’s tau-c statistic
  FREQ procedure, 158, 161
Student’s $t$ test, 384, 385
summary statistics
  insets of, 316
  saving, 278, 433
sums of squares and crossproducts, 10
superiority tests
  binomial proportions (FREQ), 172
  risk difference (FREQ), 182
$t$ test
  Student’s, 384, 385
table scores
  FREQ procedure, 152
tables
  contingency (FREQ), 70
crosstabulation (FREQ), 70, 216
multiway (FREQ), 70, 216
  one-way frequency (FREQ), 70, 214
Taratone’s adjustment
  Breslow-Day test (FREQ), 202
tests for location, 384, 462
  paired data, 384, 463
  sign test, 385
  Student’s $t$ test, 385
  Wilcoxon signed rank test, 385
tetrachoric correlation coefficient
  FREQ procedure, 164
theoretical distributions, 421
three-parameter Weibull distribution, 426
  probability plots, 426
  quantile plots, 426
threshold parameter
  probability plots, 355
quantile plots, 366
threshold parameters
  probability plots, 427
  quantile plots, 427
tiles for comparative plots
  histograms, 372
TOST
  equivalence tests (FREQ), 173, 182
trend test
  FREQ procedure, 189
trimmed means, 281, 388
two-parameter Weibull distribution, 426
  probability plots, 426
  quantile plots, 426
uncertainty coefficients
  FREQ procedure, 158, 167
univariate analysis
  for multiple variables, 439
UNIVARIATE procedure
  calculating modes, 441
  classification levels, 294
  comparative plots, 294–296, 396
  computational resources, 438
  concepts, 377
  confidence limits, 276, 386, 455
  descriptive statistics, 379, 439
  examples, 439
  extreme observations, 316, 443
  extreme values, 443
  fitted continuous distributions, 401
  frequency variables, 452
  goodness-of-fit tests, 413
  high-resolution graphics, 391
  histograms, 437, 443
  insets for descriptive statistics, 316
  keywords for insets, 316
  keywords for output data sets, 326
  line printer plots, 389, 447
  missing values, 295, 377
  mode calculation, 382
  normal probability plots, 390
  ODS graph names, 437
  ODS table names, 436
  output data sets, 326, 429, 453
  overview, 266
  percentiles, 345, 356, 382
  percentiles, confidence limits, 276, 457
  plot statements, 265
  probability plots, 345, 421
  quantile plots, 356, 421
  quantiles, confidence limits, 276
  results, 435
  robust estimates, 460
Subject Index

- robust estimators, 387
- robust location estimates, 281
- robust scale estimates, 280
- rounding, 378
- sign test, 385, 463
- specifying analysis variables, 367
- task tables, 356
- testing for location, 462
- tests for location, 384
- weight variables, 367

UNIVARIATE procedure. OUTPUT statement
output data set, 429

variances, 10

Wald confidence limits
risk difference (FREQ), 179

Weibull distribution, 413
- cdf plots, 294
deviation from theoretical distribution, 415
EDF goodness-of-fit test, 414
estimation of parameters, 315
fitting, 315, 413
formulas for, 413
histograms, 490
P-P plots, 344
probability plots, 355
quantile plots, 366, 516
three-parameter, 426
two-parameter, 426

weight values, 277
weighted kappa coefficient
FREQ procedure, 192, 194
weighted percentiles, 383
Wilcoxon signed rank test, 384, 385
Wilson confidence limits
proportions (FREQ), 169

Winsorized means, 281, 387

Yule’s Q statistic
FREQ procedure, 160

Zelen’s test
equal odds ratios (FREQ), 203
zeros, structural and random
agreement statistics (FREQ), 196
Syntax Index

AGREE option
   EXACT statement (FREQ), 87
   OUTPUT statement (FREQ), 96
   TABLES statement (FREQ), 107
   TEST statement (FREQ), 143
AJCHI option
   OUTPUT statement (FREQ), 96
ALL option
   OUTPUT statement (FREQ), 97
   PLOTS= option (CORR), 13
   PROC UNIVARIATE statement, 275
   TABLES statement (FREQ), 108
ALPHA option
   PROC CORR statement, 10
   ALPHA= option
      EXACT statement (FREQ), 91
      HISTOGRAM statement (UNIVARIATE), 304, 402, 409
      PLOTS=SCATTER option (CORR), 14
      PROBPLOT statement (UNIVARIATE), 350
      PROC UNIVARIATE statement, 275
      QQPLOT statement (UNIVARIATE), 361
      TABLES statement (FREQ), 108
   ALPHADELTA= option
      plot statements (UNIVARIATE), 369
   ALPHAINITIAL= option
      plot statements (UNIVARIATE), 369
   ANNOKEY option
      plot statements (UNIVARIATE), 369
   ANNOTATE= option
      HISTOGRAM statement (UNIVARIATE), 305, 401, 481
      PROBPLOT statement (UNIVARIATE), 350, 422
      QQPLOT statement (UNIVARIATE), 361, 422
   BETA= option
      HISTOGRAM statement (UNIVARIATE), 305, 402
      PROBPLOT statement (UNIVARIATE), 351
      QQPLOT statement (UNIVARIATE), 362
   BINOMIAL option
      EXACT statement (FREQ), 87
      OUTPUT statement (FREQ), 97
      TABLES statement (FREQ), 108
   BOWKER option
      OUTPUT statement (FREQ), 104
   BY statement
      CORR procedure, 17
      FREQ procedure, 84
      UNIVARIATE procedure, 282
   C= option
      HISTOGRAM statement (UNIVARIATE), 305, 418, 419, 488
      PROBPLOT statement (UNIVARIATE), 351
      QQPLOT statement (UNIVARIATE), 362, 516
   CAXIS= option
      plot statements (UNIVARIATE), 369
   CBARLINE= option
      HISTOGRAM statement (UNIVARIATE), 305
   CFILL= option
      INSET statement (UNIVARIATE), 324
      plot statements (UNIVARIATE), 369
   CFILLH= option
      INSET statement (UNIVARIATE), 324
   CFRAMESIDE= option
      plot statements (UNIVARIATE), 369
   CFRAMETOP= option
      plot statements (UNIVARIATE), 369
   CGRID= option
      HISTOGRAM statement (UNIVARIATE), 306
PROBPLOT statement (UNIVARIATE), 351
QQPLOT statement (UNIVARIATE), 362
CHEADER= option
INSET statement (UNIVARIATE), 324
CHISQ option
EXACT statement (FREQ), 87, 240
OUTPUT statement (FREQ), 97
TABLES statement (FREQ), 113, 240
CHREF= option
plot statements (UNIVARIATE), 370
CIBASIC option
PROC UNIVARIATE statement, 276, 455
CIPCTLDF option
PROC UNIVARIATE statement, 276
CIPCTLNORMAL option
PROC UNIVARIATE statement, 276
CIQUANTDF option
PROC UNIVARIATE statement, 457
CIQUANTNORMAL option
PROC UNIVARIATE statement, 276, 457
CL option
TABLES statement (FREQ), 116
CL= option (BINOMIAL)
TABLES statement (FREQ), 110
CL= option (RISKDIFF)
TABLES statement (FREQ), 137
CL=AGRESTICAFFO option (RISKDIFF)
TABLES statement (FREQ), 138
CL=AGRESTICOUll option (BINOMIAL)
TABLES statement (FREQ), 110
CL=ALL option (BINOMIAL)
TABLES statement (FREQ), 110
CL=ClLOPPERPEARSON option (BINOMIAL)
TABLES statement (FREQ), 110
CL=EXACT option (BINOMIAL)
TABLES statement (FREQ), 110
CL=EXACT option (RISKDIFF)
TABLES statement (FREQ), 138
CL=HA option (RISKDIFF)
TABLES statement (FREQ), 138
CL=JEFFREYS option (BINOMIAL)
TABLES statement (FREQ), 110
CL=NEWCOMBE option (RISKDIFF)
TABLES statement (FREQ), 138
CL=WALD option (BINOMIAL)
TABLES statement (FREQ), 110
CL=WALD option (RISKDIFF)
TABLES statement (FREQ), 139
CL=WILSON option (BINOMIAL)
TABLES statement (FREQ), 111
CLASS statement
UNIVARIATE procedure, 294
CLIPCURVES option
HISTOGRAM statement (UNIVARIATE), 306
CLIPREF option
HISTOGRAM statement (UNIVARIATE), 306
CMH option
OUTPUT statement (FREQ), 97
TABLES statement (FREQ), 116
CMH1 option
OUTPUT statement (FREQ), 97
TABLES statement (FREQ), 117
CMH2 option
OUTPUT statement (FREQ), 98
TABLES statement (FREQ), 117
CMHCOR option
OUTPUT statement (FREQ), 98
CMHGA option
OUTPUT statement (FREQ), 98
CMHRMS option
OUTPUT statement (FREQ), 98
COCHQ option
OUTPUT statement (FREQ), 98
COLUMN= option (RELRIISK)
EXACT statement (FREQ), 90
COLUMN= option (RISKDIFF)
EXACT statement (FREQ), 90
TABLES statement (FREQ), 139
COMMON option (RISKDIFF)
TABLES statement (FREQ), 139
COMOR option
EXACT statement (FREQ), 87
OUTPUT statement (FREQ), 100
COMPRESS option
PROC FREQ statement, 83
CONTENTS= option
HISTOGRAM statement (UNIVARIATE), 306
plot statements (UNIVARIATE), 370
TABLES statement (FREQ), 117
CONTGY option
OUTPUT statement (FREQ), 98
CONVERGE option
POLYCHORIC option (CORR), 15
POLYSERIAL option (CORR), 15
CONVERGE= option (PLCORR)
TABLES statement (FREQ), 123
CORR procedure
syntax, 9
CORR procedure, BY statement, 17
CORR procedure, FREQ statement, 17
CORR procedure, ID statement, 18
CORR procedure, PARTIAL statement, 18
CORR procedure, PLOTS= option
ALL option, 13
HISTOGRAM option, 14
MATRIX option, 13
MAXPOINTS= option, 13
NONE option, 14
NV AR= option, 14
NW ITH= option, 14, 15
SCATTER option, 14
CORR procedure, PLOTS=SCATTER option
ALPHA=, 14
ELLIPSE=, 14
NOINSET, 14
CORR procedure, POLYCHORIC
CONVERGE option, 15
MAXITER option, 15
NGROUPS option, 15
CORR procedure, POLYCHORIC option, 15
CORR procedure, POLYSERIAL
CONVERGE option, 15
MAXITER option, 16
NGROUPS option, 16
ORDINAL option, 16
CORR procedure, POLYSERIAL option, 15
CORR procedure, PROC CORR statement, 9
ALPHA option, 10
BEST= option, 10
COV option, 11
CSSCP option, 11
DATA= option, 11
EXCLNPWGT option, 11
FISHER option, 11
HOEFFDING option, 12
KENDALL option, 12
NOCORR option, 12
NOMISS option, 12
NOPRINT option, 12
NOPROB option, 12
NOSIMPLE option, 12
OUT= option, 12
OUTH= option, 12
OUTK= option, 12
OUTP= option, 12
OUTS= option, 13
PEARSON option, 13
RANK option, 16
SINGULAR= option, 16
SPEARMAN option, 16
SSCP option, 16
VARDEF= option, 16
CORR procedure, VAR statement, 18
CORR procedure, WEIGHT statement, 18
CORR procedure, WITH statement, 19
CORRECT option (BINOMIAL)
TABLES statement (FREQ), 111
CORRECT option (RISKDIFF)
TABLES statement (FREQ), 139
COV option
PROC CORR statement, 11
CPROP= option
HISTOGRAM statement (UNIVARIATE), 479
plot statements (UNIVARIATE), 370
CRAMV option
OUTPUT statement (FREQ), 98
CROSSLIST option
TABLES statement (FREQ), 117
CSHADOW= option
INSET statement (UNIVARIATE), 324
CSSCP option
PROC CORR statement, 11
CSTATREF= option
plot statements (UNIVARIATE), 370
CTEXT= option
INSET statement (UNIVARIATE), 324
plot statements (UNIVARIATE), 370
CTEXTSIDE= option
plot statements (UNIVARIATE), 370
CTEXTTOP= option
plot statements (UNIVARIATE), 370
CUMCOL option
TABLES statement (FREQ), 118
CVREF= option
plot statements (UNIVARIATE), 371
DATA option
INSET statement (UNIVARIATE), 325
DATA= option
INSET statement (UNIVARIATE), 319
PROC CORR statement, 11
PROC FREQ statement, 83
PROC UNIVARIATE statement, 277, 429
DESCENDING option
BY statement (UNIVARIATE), 282
DESCRIPTION= option
plot statements (UNIVARIATE), 371
DEVIATION option
TABLES statement (FREQ), 118
DF= option (CHISQ)
TABLES statement (FREQ), 114
ELLIPSE= option
PLOTS=SCATTER option (CORR), 14
ENDPOINTS= option
HISTOGRAM statement (UNIVARIATE), 306, 473
EQKAP option
OUTPUT statement (FREQ), 98
EQOR option
EXACT statement (FREQ), 88
OUTPUT statement (FREQ), 98
EQUAL option (RISKDIFF)
TABLES statement (FREQ), 139
EQUIVALENCE option (BINOMIAL)
TABLES statement (FREQ), 111
EQUIVALENCE option (RISKDIFF)
  TABLES statement (FREQ), 139
EQWKP option
  OUTPUT statement (FREQ), 98
EXACT option
  OUTPUT statement (FREQ), 99
  TABLES statement (FREQ), 119
EXACT statement
  FREQ procedure, 85
EXCLNPWGT option
  PROC CORR statement, 11
  PROC UNIVARIATE statement, 277
EXPECTED option
  TABLES statement (FREQ), 119
EXPONENTIAL option
  HISTOGRAM statement (UNIVARIATE), 307, 403
  PROBPLOT statement (UNIVARIATE), 351, 423
  QQPLOT statement (UNIVARIATE), 362, 423
FILL option
  HISTOGRAM statement (UNIVARIATE), 308
FISHER option
  EXACT statement (FREQ), 88
  OUTPUT statement (FREQ), 99
  PROC CORR statement, 11
  TABLES statement (FREQ), 119
FITINTERVAL= option
  plot statements (UNIVARIATE), 371
FITMETHOD= option
  plot statements (UNIVARIATE), 371
FITTOLERANCE= option
  plot statements (UNIVARIATE), 371
FONT= option
  INSET statement (UNIVARIATE), 325
  plot statements (UNIVARIATE), 371
FORMAT= option
  INSET statement (UNIVARIATE), 325
  TABLES statement (FREQ), 119
FORMCHAR= option
  PROC FREQ statement, 83
FREQ option
  PROC UNIVARIATE statement, 277, 445
FREQ procedure
  syntax, 82
FREQ procedure, BY statement, 84
FREQ procedure, EXACT statement, 85
  AGREE option, 87
  ALPHA= option, 91
  BARNARD option, 87
  BINOMIAL option, 87
  CHISQ option, 87, 240
  COLUMN= option (RELRISK), 90
  COLUMN= option (RISKDIFF), 90
  COMOR option, 87
  EQOR option, 88
  FISHER option, 88
  JT option, 88
  KAPPA option, 88
  KENTB option, 88
  LRCHI option, 88
  MAXTIME= option, 91
  MC option, 92
  MCNEM option, 88
  MEASURES option, 88
  METHOD= option (RELRISK), 90
  METHOD= option (RISKDIFF), 90
  MHCHI option, 89
  MIDP option, 92
  N= option, 92
  OR option, 89, 240
  PCHI option, 89
  PCORR option, 89
  PFORMAT= option, 92
  POINT option, 92
  RELRISK option, 89
  RISKDIFF option, 90
  SCORR option, 90
  SEED= option, 93
  SMDCR option, 90
  SMDRC option, 91
  STUTC option, 91
  TREND option, 91, 248
  WTKAP option, 91
  ZELEN option, 88
FREQ procedure, OUTPUT statement, 93
  AGREE option, 96
  AJCHI option, 96
  ALL option, 97
  BDCHI option, 97
  BINOMIAL option, 97
  BOWKER option, 104
  COCHRA option, 98
  CONTGY option, 98
  CRAMV option, 98
  EQKAP option, 98
  EQOR option, 98
  EQWKP option, 98
  EXACT option, 98
  FISHER option, 99
<table>
<thead>
<tr>
<th>Option</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAILSIMON option</td>
<td>99</td>
</tr>
<tr>
<td>GAMMA option</td>
<td>99</td>
</tr>
<tr>
<td>JT option</td>
<td>99</td>
</tr>
<tr>
<td>KAPPA option</td>
<td>99</td>
</tr>
<tr>
<td>KENTB option</td>
<td>99</td>
</tr>
<tr>
<td>LAMCR option</td>
<td>99</td>
</tr>
<tr>
<td>LAMDAS option</td>
<td>99</td>
</tr>
<tr>
<td>LAMRC option</td>
<td>99</td>
</tr>
<tr>
<td>LGOR option</td>
<td>100</td>
</tr>
<tr>
<td>LGRRC1 option</td>
<td>100</td>
</tr>
<tr>
<td>LGRRC2 option</td>
<td>100</td>
</tr>
<tr>
<td>LRCHI option</td>
<td>100</td>
</tr>
<tr>
<td>MCNEM option</td>
<td>100</td>
</tr>
<tr>
<td>MEASURES option</td>
<td>100</td>
</tr>
<tr>
<td>MHCHI option</td>
<td>100</td>
</tr>
<tr>
<td>MHOR option</td>
<td>100</td>
</tr>
<tr>
<td>MHRRC1 option</td>
<td>100</td>
</tr>
<tr>
<td>MHRRC2 option</td>
<td>101</td>
</tr>
<tr>
<td>N option</td>
<td>101</td>
</tr>
<tr>
<td>NMISS option</td>
<td>101</td>
</tr>
<tr>
<td>OR option</td>
<td>101</td>
</tr>
<tr>
<td>OUT= option</td>
<td>93</td>
</tr>
<tr>
<td>PCHI option</td>
<td>101</td>
</tr>
<tr>
<td>PCORR option</td>
<td>101</td>
</tr>
<tr>
<td>PLCORR option</td>
<td>101</td>
</tr>
<tr>
<td>RDIF1 option</td>
<td>101</td>
</tr>
<tr>
<td>RDIF2 option</td>
<td>102</td>
</tr>
<tr>
<td>RELRISK option</td>
<td>102</td>
</tr>
<tr>
<td>RISK1 option</td>
<td>102</td>
</tr>
<tr>
<td>RISK11 option</td>
<td>103</td>
</tr>
<tr>
<td>RISK12 option</td>
<td>103</td>
</tr>
<tr>
<td>RISK2 option</td>
<td>103</td>
</tr>
<tr>
<td>RISK21 option</td>
<td>103</td>
</tr>
<tr>
<td>RISK22 option</td>
<td>103</td>
</tr>
<tr>
<td>RISKDIFF option</td>
<td>102</td>
</tr>
<tr>
<td>RISKDIFF1 option</td>
<td>102</td>
</tr>
<tr>
<td>RISKDIFF2 option</td>
<td>102</td>
</tr>
<tr>
<td>RRC1 option</td>
<td>102</td>
</tr>
<tr>
<td>RRC2 option</td>
<td>102</td>
</tr>
<tr>
<td>SCORR option</td>
<td>103</td>
</tr>
<tr>
<td>SMDCR option</td>
<td>103</td>
</tr>
<tr>
<td>SMDRC option</td>
<td>103</td>
</tr>
<tr>
<td>STUTC option</td>
<td>103</td>
</tr>
<tr>
<td>TAUB option</td>
<td>99</td>
</tr>
<tr>
<td>TAUC option</td>
<td>103</td>
</tr>
<tr>
<td>TREND option</td>
<td>104</td>
</tr>
<tr>
<td>TSYM option</td>
<td>104</td>
</tr>
<tr>
<td>U option</td>
<td>104</td>
</tr>
<tr>
<td>UCR option</td>
<td>104</td>
</tr>
<tr>
<td>URC option</td>
<td>104</td>
</tr>
<tr>
<td>WTKAP option</td>
<td>104</td>
</tr>
<tr>
<td>ZELEN option</td>
<td>98</td>
</tr>
</tbody>
</table>

**FREQ procedure, PROC FREQ statement, 82**

<table>
<thead>
<tr>
<th>Option</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPRESS option</td>
<td>83</td>
</tr>
<tr>
<td>DATA= option</td>
<td>83</td>
</tr>
<tr>
<td>FORMCHAR= option</td>
<td>83</td>
</tr>
<tr>
<td>NLEVELS option</td>
<td>84</td>
</tr>
<tr>
<td>NOPRINT option</td>
<td>84</td>
</tr>
<tr>
<td>ORDER= option</td>
<td>84</td>
</tr>
<tr>
<td>PAGE option</td>
<td>84</td>
</tr>
<tr>
<td>AGREE option</td>
<td>107</td>
</tr>
<tr>
<td>ALL option</td>
<td>108</td>
</tr>
<tr>
<td>ALPHA= option</td>
<td>108</td>
</tr>
<tr>
<td>BDT option (CMH)</td>
<td>116</td>
</tr>
<tr>
<td>BINOMIAL option</td>
<td>108</td>
</tr>
<tr>
<td>CELLCHI2 option</td>
<td>112</td>
</tr>
<tr>
<td>CHISQ option</td>
<td>113</td>
</tr>
<tr>
<td>CMH option</td>
<td>116</td>
</tr>
<tr>
<td>CL option</td>
<td>116</td>
</tr>
<tr>
<td>CL= option (BINOMIAL)</td>
<td>110</td>
</tr>
<tr>
<td>CL= option (RISKDIFF)</td>
<td>137</td>
</tr>
<tr>
<td>CL= option (RISKDIFF)</td>
<td>138</td>
</tr>
<tr>
<td>CL=AGRESTICAFFO option</td>
<td>138</td>
</tr>
<tr>
<td>CL=AGRESTICOULL option</td>
<td>110</td>
</tr>
<tr>
<td>CL=ALL option (BINOMIAL)</td>
<td>110</td>
</tr>
<tr>
<td>CL=CLOPPERPEARSON option</td>
<td>110</td>
</tr>
<tr>
<td>CL=EXACT option (BINOMIAL)</td>
<td>110</td>
</tr>
<tr>
<td>CL=EXACT option (RISKDIFF)</td>
<td>138</td>
</tr>
<tr>
<td>CL=HA option (RISKDIFF)</td>
<td>138</td>
</tr>
<tr>
<td>CL=JEFFREYS option (BINOMIAL)</td>
<td>110</td>
</tr>
<tr>
<td>CL=NEWCOMBE option (RISKDIFF)</td>
<td>138</td>
</tr>
<tr>
<td>CL=WALD option (BINOMIAL)</td>
<td>110</td>
</tr>
<tr>
<td>CL=WALD option (RISKDIFF)</td>
<td>139</td>
</tr>
<tr>
<td>CL=WILSON option (BINOMIAL)</td>
<td>111</td>
</tr>
<tr>
<td>CMH option</td>
<td>116</td>
</tr>
<tr>
<td>CMH1 option</td>
<td>117</td>
</tr>
<tr>
<td>CMH2 option</td>
<td>117</td>
</tr>
<tr>
<td>COLUMN= option</td>
<td>139</td>
</tr>
<tr>
<td>COMMON option</td>
<td>139</td>
</tr>
<tr>
<td>CONTENTS= option</td>
<td>117</td>
</tr>
<tr>
<td>CONVERGE= option (PLCORR)</td>
<td>123</td>
</tr>
<tr>
<td>CORRECT option</td>
<td>111</td>
</tr>
<tr>
<td>CORRECT option (RISKDIFF)</td>
<td>139</td>
</tr>
<tr>
<td>CROSSLIST option</td>
<td>117</td>
</tr>
<tr>
<td>CUMCOL option</td>
<td>118</td>
</tr>
<tr>
<td>DEVIATION option</td>
<td>118</td>
</tr>
<tr>
<td>DF= option (CHISQ)</td>
<td>114</td>
</tr>
<tr>
<td>EQUAL option (RISKDIFF)</td>
<td>139</td>
</tr>
<tr>
<td>EQUIVALENCE option (BINOMIAL)</td>
<td>111</td>
</tr>
<tr>
<td>EQUIVALENCE option (RISKDIFF)</td>
<td>139</td>
</tr>
<tr>
<td>EXACT option</td>
<td>119</td>
</tr>
<tr>
<td>EXPECTED option</td>
<td>119</td>
</tr>
<tr>
<td>FISHER option</td>
<td>119</td>
</tr>
<tr>
<td>FORMAT= option</td>
<td>119</td>
</tr>
<tr>
<td>GAILSIMON option</td>
<td>119</td>
</tr>
<tr>
<td>JT option</td>
<td>119</td>
</tr>
<tr>
<td>LEVEL= option (BINOMIAL)</td>
<td>111</td>
</tr>
</tbody>
</table>
LIST option, 120
LRCHI option (CHISQ), 114
MANTELLEFISS option (CMH), 116
MARGIN= option (BINOMIAL), 111
MARGIN= option (RISKDIFF), 139
MAXITER= option (PLCORR), 124
MAXLEVELS= option, 120
MEASURES option, 120
METHOD= option (RISKDIFF), 140
METHOD=FM option (RISKDIFF), 140
METHOD=HA option (RISKDIFF), 140
METHOD=NEWCOMBE option (RISKDIFF), 140
METHOD=WALD option (RISKDIFF), 140
MISSING option, 120
MISSPRINT option, 121
NOCOL option, 121
NOCUM option, 121
NOFREQ option, 121
NONINFERIORITY option (BINOMIAL), 112
NONINFERIORITY option (RISKDIFF), 140
NOPERCENT option, 121
NOPRINT option, 121
NORISKS option (RISKDIFF), 141
NOROW option, 121
NOWARN option, 122
OR option, 122, 136
OUT= option, 123
OUTCUM option, 123
OUTEXPECT option, 123, 228
OUTPCT option, 123
P= option (BINOMIAL), 112
PEARSONRES option (CROSSTLST), 118
PLCORR option, 123
PLOTS= option, 124
PLOTS=AGREEPLOT option, 125
PLOTS=CUMFREQPLOT option, 126
PLOTS=DEVIATIONPLOT option, 126
PLOTS=FREQPLOT option, 126
PLOTS=KAPPAPLOT option, 127
PLOTS=MOSAICPLOT option, 128
PLOTS=NONE option, 129
PLOTS=ODDSRATIOPLOT option, 129
PLOTS=RELRISKPLOT option, 129
PLOTS=RISKDIFFPLOT option, 130
PLOTS=WTKAPPAPLOT, 130
PRNTKWT option, 136
RELRIK option, 136, 240
RISKDIFF option, 136
SCORES= option, 141, 251
SCOROUT option, 141
SPARSE option, 142, 228
STDRES option (CROSSTLST), 118
SUPERIORITY option (BINOMIAL), 112
SUPERIORITY option (RISKDIFF), 141
TESTF= option, 154
TESTF= option (CHISQ), 114
TESTP= option, 154, 234
TESTP= option (CHISQ), 115
TOTPCT option, 142
TREND option, 142, 248
VAR= option (BINOMIAL), 112
VAR= option (RISKDIFF), 141
WARN= option (CHISQ), 115
FREQ procedure, TEST statement, 142
AGREE option, 143
GAMMA option, 143
KAPPA option, 144
MEASURES option, 144
PCORR option, 144
PLCORR option, 144
SCORR option, 144
SMDCR option, 145, 248
SMDCR option, 145
STUTC option, 145
TAUB option, 144
TAUC option, 145
WTKAP option, 145
FREQ procedure, WEIGHT statement, 145
ZEROS option, 146
FREQ statement
CORR procedure, 17
UNIVARIATE procedure, 296
FRONTREF option
HISTOGRAM statement (UNIVARIATE), 308
GAILSIMON option
OUTPUT statement (FREQ), 99
TABLES statement (FREQ), 119
GAMMA option
HISTOGRAM statement (UNIVARIATE), 308, 403, 410, 411, 483
OUTPUT statement (FREQ), 99
PROBPLOT statement (UNIVARIATE), 351, 423
QQPLOT statement (UNIVARIATE), 362, 423
TEST statement (FREQ), 143
GOUT= option
PROC UNIVARIATE statement, 277
GRID option
HISTOGRAM statement (UNIVARIATE), 308
PROBPLOT statement (UNIVARIATE), 352
QQPLOT statement (UNIVARIATE), 362, 364, 512
GUMBEL option
HISTOGRAM statement (UNIVARIATE), 308, 404
PROBPLOT statement (UNIVARIATE), 352
QQPLOT statement (UNIVARIATE), 363

Gumbel option
PROBPLOT statement (UNIVARIATE), 424
QQPLOT statement (UNIVARIATE), 424

GUTTER= option
INSET statement (UNIVARIATE), 325

HAXIS= option
plot statements (UNIVARIATE), 371
HEADER= option
INSET statement (UNIVARIATE), 325
HEIGHT= option
INSET statement (UNIVARIATE), 325
plot statements (UNIVARIATE), 371

HISTOGRAM
PLOTS= option (CORR), 14

HISTOGRAM statement
UNIVARIATE procedure, 296

HMINOR= option
plot statements (UNIVARIATE), 371

HOEFFDING option
PROC CORR statement, 12

HOFFSET= option
HISTOGRAM statement (UNIVARIATE), 309

HREF= option
plot statements (UNIVARIATE), 372
HREFLABELS= option
plot statements (UNIVARIATE), 372
HREFLABPOS= option
plot statements (UNIVARIATE), 372

ID statement
CORR procedure, 18
UNIVARIATE procedure, 316

IDOUT option
PROC UNIVARIATE statement, 277

IGAUSS option
HISTOGRAM statement (UNIVARIATE), 309, 405

INFONT= option
plot statements (UNIVARIATE), 372

INHEIGHT= option
plot statements (UNIVARIATE), 372

INSET statement
UNIVARIATE procedure, 316

INTERBAR= option
HISTOGRAM statement (UNIVARIATE), 310

INTERTILE= option
HISTOGRAM statement (UNIVARIATE), 479
plot statements (UNIVARIATE), 372

ITPRINT option
plot statements (UNIVARIATE), 372

JT option
EXACT statement (FREQ), 88
OUTPUT statement (FREQ), 99
TABLES statement (FREQ), 119

K= option
HISTOGRAM statement (UNIVARIATE), 310, 418, 419

KAPPA option
EXACT statement (FREQ), 88
OUTPUT statement (FREQ), 99
TEST statement (FREQ), 144

KENDALL option
PROC CORR statement, 12

KENTB option
EXACT statement (FREQ), 88
OUTPUT statement (FREQ), 99
TEST statement (FREQ), 144

KERNEL option
HISTOGRAM statement (UNIVARIATE), 310, 418, 419, 488

KEYLEVEL= option
CLASS statement (UNIVARIATE), 296
PROC UNIVARIATE statement, 469

L= option
plot statements (UNIVARIATE), 373

LABEL= option
QQPLOT statement (UNIVARIATE), 364

LAMBDA= option
HISTOGRAM statement (UNIVARIATE), 310
PPPLOT statement (UNIVARIATE), 340

LAMCR option
OUTPUT statement (FREQ), 99

LAMDA option
OUTPUT statement (FREQ), 99

LAMRC option
OUTPUT statement (FREQ), 99

LEVEL= option (BINOMIAL)
TABLES statement (FREQ), 111

LGOR option
OUTPUT statement (FREQ), 100

LGRID= option
HISTOGRAM statement (UNIVARIATE), 310
PROBPLOT statement (UNIVARIATE), 352
QQPLOT statement (UNIVARIATE), 363, 364

LGRRC1 option
OUTPUT statement (FREQ), 100

LGRRC2 option
OUTPUT statement (FREQ), 100

LHREF= option
plot statements (UNIVARIATE), 373

LIST option
TABLES statement (FREQ), 120

LOCCOUNT option
PROC UNIVARIATE statement, 277, 462
LOGNORMAL option
   HISTOGRAM statement (UNIVARIATE), 310, 406, 483, 489, 501
   PROBPLOT statement (UNIVARIATE), 352, 424, 497
   QQPLOT statement (UNIVARIATE), 363, 424
LOWER= option
   HISTOGRAM statement (UNIVARIATE), 310
LRCHI option
   EXACT statement (FREQ), 88
   OUTPUT statement (FREQ), 100
LRCHI option (CHISQ)
   TABLES statement (FREQ), 114
LSTATREF= option
   plot statements (UNIVARIATE), 373
LVREF= option
   plot statements (UNIVARIATE), 373
MANTELFLEISS option (CMH)
   TABLES statement (FREQ), 116
MARGIN= option (BINOMIAL)
   TABLES statement (FREQ), 111
MARGIN= option (RISKDIFF)
   TABLES statement (FREQ), 139
MATRIX option
   PLOTS= option (CORR), 13
MAXITER option
   POLYCHORIC option (CORR), 15
   POLYSERIAL option (CORR), 16
MAXITER= option
   plot statements (UNIVARIATE), 373
MAXITER= option (PLCORR)
   TABLES statement (FREQ), 124
MAXLEVELS= option
   TABLES statement (FREQ), 120
MAXBIN= option
   HISTOGRAM statement (UNIVARIATE), 310
MAXPOINTS= option
   PLOTS= option (CORR), 13
MAXSIGMAS= option
   HISTOGRAM statement (UNIVARIATE), 311
MAXTIME= option
   EXACT statement (FREQ), 91
MC option
   EXACT statement (FREQ), 92
MCNEM option
   EXACT statement (FREQ), 88
   OUTPUT statement (FREQ), 100
MEASURES option
   EXACT statement (FREQ), 88
   OUTPUT statement (FREQ), 100
   TABLES statement (FREQ), 120
   TEST statement (FREQ), 144
METHOD= option (RELRISK)
   EXACT statement (FREQ), 90
METHOD= option (RISKDIFF)
   EXACT statement (FREQ), 90
   TABLES statement (FREQ), 140
METHOD=FM option (RISKDIFF)
   TABLES statement (FREQ), 140
METHOD=HA option (RISKDIFF)
   TABLES statement (FREQ), 140
METHOD=NEWCOMBE option (RISKDIFF)
   TABLES statement (FREQ), 140
METHOD=WALD option (RISKDIFF)
   TABLES statement (FREQ), 140
MHCHI option
   EXACT statement (FREQ), 89
   OUTPUT statement (FREQ), 100
MHR option
   OUTPUT statement (FREQ), 100
MHRRC1 option
   OUTPUT statement (FREQ), 100
MHRRC2 option
   OUTPUT statement (FREQ), 101
MIDP option
   EXACT statement (NPAR1WAY), 92
MIDPERCENTS option
   HISTOGRAM statement (UNIVARIATE), 311, 476
MIDPOINTS= option
   HISTOGRAM statement (UNIVARIATE), 311, 471, 473
MISSING option
   CLASS statement (UNIVARIATE), 295
   TABLES statement (FREQ), 120
MISSPRINT option
   TABLES statement (FREQ), 121
MODES option
   PROC UNIVARIATE statement, 277, 441
MU= option
   PROC UNIVARIATE statement, 277
MU= option
   HISTOGRAM statement (UNIVARIATE), 312, 476
   PROBPLOT statement (UNIVARIATE), 352
   QQPLOT statement (UNIVARIATE), 363, 505
N option
   OUTPUT statement (FREQ), 101
N= option
   EXACT statement (FREQ), 92
NADJ= option
   PROBPLOT statement (UNIVARIATE), 352
   QQPLOT statement (UNIVARIATE), 363, 420
NAME= option
   plot statements (UNIVARIATE), 373
NCOLS= option
  INSET statement (UNIVARIATE), 325
  plot statements (UNIVARIATE), 373
NENDPOINTS= option
  HISTOGRAM statement (UNIVARIATE), 312
NEXTROBS= option
  PROC UNIVARIATE statement, 278, 443
NEXTRVAL= option
  PROC UNIVARIATE statement, 278, 443
NGROUPS option
  POLYCHORIC option (CORR), 15
  POLYSERIAL option (CORR), 16
NLEVELS option
  PROC FREQ statement, 84
NMIDPOINTS= option
  HISTOGRAM statement (UNIVARIATE), 312
NMISS option
  OUTPUT statement (FREQ), 101
NOBARS option
  HISTOGRAM statement (UNIVARIATE), 312
NOFRAME option
  INSET statement (UNIVARIATE), 325
  plot statements (UNIVARIATE), 373
NOFREQ option
  TABLES statement (FREQ), 121
NOCOL option
  TABLES statement (FREQ), 121
NOCORR option
  PROC CORR statement, 12
NOCUM option
  TABLES statement (FREQ), 121
NOFRAME option
  INSET statement (UNIVARIATE), 325
  plot statements (UNIVARIATE), 373
NOFREQ option
  TABLES statement (FREQ), 121
NOHLABEL option
  plot statements (UNIVARIATE), 373
NOINSET option
  PLOTS=SCATTER option (CORR), 14
NOKEYMOVE option
  CLASS statement (UNIVARIATE), 296
NOMISS option
  PROC CORR statement, 12
NONE option
  PLOTS= option (CORR), 14
NONINTEROORDITY option (BINOMIAL)
  TABLES statement (FREQ), 112
NONINTEROORDITY option (RISKDIFF)
  TABLES statement (FREQ), 140
NOPERCENT option
  TABLES statement (FREQ), 121
NOPLOT option
  HISTOGRAM statement (UNIVARIATE), 312
NOPRINT option
  HISTOGRAM statement (UNIVARIATE), 312
  PROC CORR statement, 12
  PROC FREQ statement, 84
  PROC UNIVARIATE statement, 278
  TABLES statement (FREQ), 121
NOPROB option
  PROC CORR statement, 12
NORISKS option (RISKDIFF)
  TABLES statement (FREQ), 141
NORMAL option
  HISTOGRAM statement (UNIVARIATE), 312, 406, 476
  PROBPLOT statement (UNIVARIATE), 352, 424
  PROC UNIVARIATE statement, 278
  QQPLOT statement (UNIVARIATE), 363, 424
NORMALTEST option
  PROC UNIVARIATE statement, 278
NOROW option
  TABLES statement (FREQ), 121
NOSIMPLE option
  PROC CORR statement, 12
NOSPARSE option
  TABLES statement (FREQ), 121
NOTABCONTENTS option
  HISTOGRAM statement (UNIVARIATE), 312
  PROC UNIVARIATE statement, 278
NOTSORTED option
  BY statement (UNIVARIATE), 282
NOVARCONTENTS option
  PROC UNIVARIATE statement, 278
NOVLABEL option
  plot statements (UNIVARIATE), 374
NOVICK option
  plot statements (UNIVARIATE), 374
NOWARN option
  TABLES statement (FREQ), 122
NROWS= option
  TABLES statement (FREQ), 122
NWITH= option
  PLOTS= option (CORR), 14
NOWITH= option
  PLOTS= option (CORR), 14, 15
OR option
  EXACT statement (FREQ), 89, 240
  OUTPUT statement (FREQ), 101
  TABLES statement (FREQ), 122, 136
ORDER= option
  CLASS statement (UNIVARIATE), 295
  PROC FREQ statement, 84
ORDINAL option
  POLYSERIAL option (CORR), 16
OUT= option
  OUTPUT statement (FREQ), 93
  OUTPUT statement (UNIVARIATE), 326
### Syntax Index

<table>
<thead>
<tr>
<th>Statement</th>
<th>Procedure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROC CORR statement</td>
<td>12</td>
<td>PROC UNIVARIATE statement, 278, 382</td>
</tr>
<tr>
<td>TABLES statement (FREQ)</td>
<td>12</td>
<td>PCTLGROUP= option</td>
</tr>
<tr>
<td>OUTCUM option</td>
<td>123</td>
<td>OUTPUT statement (UNIVARIATE), 329</td>
</tr>
<tr>
<td>TABLES statement (FREQ)</td>
<td>123</td>
<td>PCTLMINOR option</td>
</tr>
<tr>
<td>OUTEXPECT option</td>
<td>123</td>
<td>PROBPLOT statement (UNIVARIATE), 353</td>
</tr>
<tr>
<td>OUTH= option</td>
<td>123, 228</td>
<td>QQPLOT statement (UNIVARIATE), 364</td>
</tr>
<tr>
<td>PROC CORR statement</td>
<td>12</td>
<td>PCTLNAME= option</td>
</tr>
<tr>
<td>OUTHISTOGRAM= option</td>
<td>123</td>
<td>OUTPUT statement (UNIVARIATE), 330</td>
</tr>
<tr>
<td>HISTOGRAM statement (UNIVARIATE), 313, 431, 473</td>
<td>123</td>
<td>PCTLORDER= option</td>
</tr>
<tr>
<td>OUTK= option</td>
<td>12</td>
<td>PROBPLOT statement (UNIVARIATE), 353</td>
</tr>
<tr>
<td>PROC CORR statement</td>
<td>12</td>
<td>PCTLPRE= option</td>
</tr>
<tr>
<td>OUTKERNEL= option</td>
<td>12</td>
<td>OUTPUT statement (UNIVARIATE), 331</td>
</tr>
<tr>
<td>HISTOGRAM statement (UNIVARIATE), 313, 432</td>
<td>12</td>
<td>PCTLNNAME= option</td>
</tr>
<tr>
<td>OUTP= option</td>
<td>12</td>
<td>OUTPUT statement (UNIVARIATE), 331</td>
</tr>
<tr>
<td>PROC CORR statement</td>
<td>12</td>
<td>PCTLSCALE option</td>
</tr>
<tr>
<td>OUTPUT statement</td>
<td>12</td>
<td>QQPLOT statement (UNIVARIATE), 364, 428</td>
</tr>
<tr>
<td>FREQ procedure, 93</td>
<td>12</td>
<td>PEARSON option</td>
</tr>
<tr>
<td>UNIVARIATE procedure, 326, 367</td>
<td>12</td>
<td>PROC CORR statement, 13</td>
</tr>
<tr>
<td>OUTS= option</td>
<td>13</td>
<td>PEARSONRES option (CROSSLIST)</td>
</tr>
<tr>
<td>PROC CORR statement</td>
<td>13</td>
<td>TABLES statement (FREQ), 118</td>
</tr>
<tr>
<td>OUTTABLE= option</td>
<td>12</td>
<td>PERCENTS= option</td>
</tr>
<tr>
<td>PROC UNIVARIATE statement, 278, 433</td>
<td>12</td>
<td>HISTOGRAM statement (UNIVARIATE), 313</td>
</tr>
<tr>
<td>OVERLAY option</td>
<td>12</td>
<td>PFORMAT= option</td>
</tr>
<tr>
<td>plot statements (UNIVARIATE), 375</td>
<td>12</td>
<td>EXACT statement (FREQ), 92</td>
</tr>
<tr>
<td>P= option (BINOMIAL)</td>
<td>112</td>
<td>PLCORR option</td>
</tr>
<tr>
<td>TABLES statement (FREQ)</td>
<td>112</td>
<td>OUTPUT statement (FREQ), 101</td>
</tr>
<tr>
<td>PAGE option</td>
<td>84</td>
<td>TABLES statement (FREQ), 123</td>
</tr>
<tr>
<td>PROC FREQ statement</td>
<td>84</td>
<td>TEST statement (FREQ), 144</td>
</tr>
<tr>
<td>PAGE option</td>
<td>84</td>
<td>PLOT option</td>
</tr>
<tr>
<td>PROC FREQ statement</td>
<td>84</td>
<td>PROC UNIVARIATE statement, 447</td>
</tr>
<tr>
<td>PARETO option</td>
<td>407</td>
<td>plot statements</td>
</tr>
<tr>
<td>HISTOGRAM statement (UNIVARIATE), 313, 407</td>
<td>447</td>
<td>UNIVARIATE procedure, 369</td>
</tr>
<tr>
<td>PROBPLOT statement (UNIVARIATE), 353</td>
<td>447</td>
<td>PLOTS option</td>
</tr>
<tr>
<td>QQPLOT statement (UNIVARIATE), 363</td>
<td>447</td>
<td>PROC UNIVARIATE statement, 278</td>
</tr>
<tr>
<td>Pareto option</td>
<td>425</td>
<td>PLOTS option</td>
</tr>
<tr>
<td>PROBPLOT statement (UNIVARIATE), 425</td>
<td>425</td>
<td>TABLES statement (FREQ), 124</td>
</tr>
<tr>
<td>QQPLOT statement (UNIVARIATE), 425</td>
<td>425</td>
<td>PLOTS= option</td>
</tr>
<tr>
<td>PARTIAL statement</td>
<td>18</td>
<td>TABLES statement (FREQ), 124</td>
</tr>
<tr>
<td>CORR procedure, 18</td>
<td>18</td>
<td>TABLES statement (FREQ), 125</td>
</tr>
<tr>
<td>PCHI option</td>
<td>89</td>
<td>PLOTS=AGREEPLOT option</td>
</tr>
<tr>
<td>EXACT statement (FREQ), 89</td>
<td>89</td>
<td>TABLES statement (FREQ), 125</td>
</tr>
<tr>
<td>OUTPUT statement (FREQ), 101</td>
<td>101</td>
<td>PLOTS=CUMFREQPLOT option</td>
</tr>
<tr>
<td>PCORR option</td>
<td>89</td>
<td>TABLES statement (FREQ), 126</td>
</tr>
<tr>
<td>EXACT statement (FREQ), 89</td>
<td>89</td>
<td>PLOTS=DEVIATIONPLOT option</td>
</tr>
<tr>
<td>OUTPUT statement (FREQ), 101</td>
<td>101</td>
<td>TABLES statement (FREQ), 126</td>
</tr>
<tr>
<td>TEST statement (FREQ), 144</td>
<td>144</td>
<td>PLOTS=FREQPLOT option</td>
</tr>
<tr>
<td>PCTLAXIS option</td>
<td>144</td>
<td>TABLES statement (FREQ), 126</td>
</tr>
<tr>
<td>QQPLOT statement (UNIVARIATE), 364, 428, 512</td>
<td>144</td>
<td>PLOTS=KAPPAPLOT option</td>
</tr>
<tr>
<td>PCTLDEF= option</td>
<td>144</td>
<td>TABLES statement (FREQ), 127</td>
</tr>
<tr>
<td>PCTLNAME= option</td>
<td>144</td>
<td>PLOTS=MOSAICPLOT option</td>
</tr>
<tr>
<td>PCTLENOR= option</td>
<td>144</td>
<td>TABLES statement (FREQ), 128</td>
</tr>
<tr>
<td>PCTLORDER= option</td>
<td>144</td>
<td>PLOTS=NONE option</td>
</tr>
<tr>
<td>PCTLPRE= option</td>
<td>144</td>
<td>TABLES statement (FREQ), 129</td>
</tr>
</tbody>
</table>
Syntax Index

PLOTS=ODDSRATIO option
   TABLES statement (FREQ), 129
PLOTS=RELRISKPLOT option
   TABLES statement (FREQ), 129
PLOTS=RISKDIFFPLOT option
   TABLES statement (FREQ), 130
PLOTS=WTKAPPAPLOT option
   TABLES statement (FREQ), 130
PLOTSIZE= option
   PROC UNIVARIATE statement, 280
POINT option
   EXACT statement (FREQ), 92
POSITION= option
   INSET statement (FREQ), 325
POWER option
   HISTOGRAM statement (UNIVARIATE), 313, 408
   PROBPLOT statement (UNIVARIATE), 353
   QQPLOT statement (UNIVARIATE), 364
power option
   PROBPLOT statement (UNIVARIATE), 425
   QQPLOT statement (UNIVARIATE), 425
PPPLOT statement
   options dictionary, 337
   options summarized by function, 335, 336
   UNIVARIATE procedure, 332
PRINTKWTS option
   TABLES statement (FREQ), 108, 136
PROBPLOT statement
   UNIVARIATE procedure, 345
PROC CORR statement, 9, see CORR procedure
   CORR procedure, 9
PROC FREQ statement, see FREQ procedure
PROC UNIVARIATE statement, 274, see UNIVARIATE procedure
QQPLOT statement
   UNIVARIATE procedure, 356
RANK option
   PROC CORR statement, 16
RANKADJ= option
   PROBPLOT statement (UNIVARIATE), 353
   QQPLOT statement (UNIVARIATE), 365, 420
RAYLEIGH option
   HISTOGRAM statement (UNIVARIATE), 314, 409
   PROBPLOT statement (UNIVARIATE), 353
   QQPLOT statement (UNIVARIATE), 364
Rayleigh option
   PROBPLOT statement (UNIVARIATE), 425
   QQPLOT statement (UNIVARIATE), 425
RDIF1 option
   OUTPUT statement (FREQ), 101
RDIF2 option
   OUTPUT statement (FREQ), 102
REFPOINT= option
   INSET statement (UNIVARIATE), 325
RELRISK option
   EXACT statement (FREQ), 89
   OUTPUT statement (FREQ), 102
   TABLES statement (FREQ), 136, 240
RISK1 option
   OUTPUT statement (FREQ), 102
RISK11 option
   OUTPUT statement (FREQ), 103
RISK12 option
   OUTPUT statement (FREQ), 103
RISK2 option
   OUTPUT statement (FREQ), 103
RISK21 option
   OUTPUT statement (FREQ), 103
RISK22 option
   OUTPUT statement (FREQ), 103
RISKDIFF option
   EXACT statement (FREQ), 90
   OUTPUT statement (FREQ), 102
   TABLES statement (FREQ), 136
RISKDIFF1 option
   OUTPUT statement (FREQ), 102
RISKDIFF2 option
   OUTPUT statement (FREQ), 102
ROBSTSCALE option
   PROC UNIVARIATE statement, 280, 460
ROTATE option
   PROBPLOT statement (UNIVARIATE), 353
   QQPLOT statement (UNIVARIATE), 365
ROUND= option
   PROC UNIVARIATE statement, 280
RRC1 option
   OUTPUT statement (FREQ), 102
RRC2 option
   OUTPUT statement (FREQ), 102
RTINCLUDE option
   HISTOGRAM statement (UNIVARIATE), 314, 473
SB option
   HISTOGRAM statement (UNIVARIATE), 314
SCALE= option
   HISTOGRAM statement (UNIVARIATE), 403–405, 481
SCATTER option
   PLOTS= option (CORR), 14
SCORES= option
   TABLES statement (FREQ), 141, 251
SCOROUT option
   TABLES statement (FREQ), 141
SCORR option
   EXACT statement (FREQ), 90
   OUTPUT statement (FREQ), 103
   TEST statement (FREQ), 144
SEED= option
   EXACT statement (FREQ), 93
SIGMA= option
   HISTOGRAM statement (UNIVARIATE), 314, 402, 409, 476
   PROBPLOT statement (UNIVARIATE), 354, 497
   QQPLOT statement (UNIVARIATE), 365, 505, 508
SINGULAR= option
   PROC CORR statement, 16
SLOPE= option
   PROBPLOT statement (UNIVARIATE), 354
   QQPLOT statement (UNIVARIATE), 365
SMDCR option
   EXACT statement (FREQ), 90
   OUTPUT statement (FREQ), 103
   TEST statement (FREQ), 145, 248
SMDRC option
   EXACT statement (FREQ), 91
   OUTPUT statement (FREQ), 103
   TEST statement (FREQ), 145
SPARSE option
   TABLES statement (FREQ), 142, 228
SPEARMAN option
   PROC CORR statement, 16
SQUARE option
   PROBPLOT statement (UNIVARIATE), 355, 497
   QQPLOT statement (UNIVARIATE), 505
   QQPLOT statement (UNIVARIATE), 366
SSCP option
   PROC CORR statement, 16
STATREFLABELS= option
   UNIVARIATE procedure, 376
STDRES option (CROSSLIST)
   TABLES statement (FREQ), 118
STUTC option
   EXACT statement (FREQ), 91
   OUTPUT statement (FREQ), 103
   TEST statement (FREQ), 145
SU option
   HISTOGRAM statement (UNIVARIATE), 315
SUMMARYCONTENTS= option
   PROC UNIVARIATE statement, 280
SUPERIORITY option (BINOMIAL)
   TABLES statement (FREQ), 112
SUPERIORITY option (RISKDIFF)
   TABLES statement (FREQ), 141
TABLES statement
   FREQ procedure, 104
TAUB option
   OUTPUT statement (FREQ), 99
   TEST statement (FREQ), 144
TAUC option
   OUTPUT statement (FREQ), 103
   TEST statement (FREQ), 145
TEST statement
   FREQ procedure, 142
TESTF= option
   TABLES statement (FREQ), 154
TESTF= option (CHISQ)
   TABLES statement (FREQ), 114
TESTP= option
   TABLES statement (FREQ), 154, 234
TESTP= option (CHISQ)
   TABLES statement (FREQ), 115
THETA= option
   HISTOGRAM statement (UNIVARIATE), 315, 402, 409, 481, 489, 501
   PROBPLOT statement (UNIVARIATE), 355
   QQPLOT statement (UNIVARIATE), 366
THRESHOLD= option
   HISTOGRAM statement (UNIVARIATE), 315, 404
   PROBPLOT statement (UNIVARIATE), 355
   QQPLOT statement (UNIVARIATE), 366
TOTPCT option
   TABLES statement (FREQ), 142
TREND option
   EXACT statement (FREQ), 91, 248
   OUTPUT statement (FREQ), 104
   TABLES statement (FREQ), 142, 248
TRIMMED= option
   PROC UNIVARIATE statement, 281, 460
TSYMM option
   OUTPUT statement (FREQ), 104
TURNVLABELS option
   plot statements (UNIVARIATE), 376
U option
   OUTPUT statement (FREQ), 104
UCR option
   OUTPUT statement (FREQ), 104
UNIVARIATE procedure
   syntax, 274
   UNIVARIATE procedure, BY statement, 282
   DESCENDING option, 282
   NOTSORTED option, 282
   UNIVARIATE procedure, CDFPLOT statement, 282
   ALPHA= beta-option, 288
   ALPHA= gamma-option, 288
   BETA beta-option, 288
   BETA= option, 288
   C= option, 289
EXPONENTIAL option, 289
GAMMA option, 289
GAMMA= option, 308
GUMBEL option, 290
IGAUSS option, 290
LAMBDA= option, 290
LOGNORMAL option, 291
MU= option, 291
NOECDF option, 291
NORMAL option, 291
PARETO option, 292
POWER option, 292
RAYLEIGH option, 293
SIGMA= option, 293
THETA= option, 293
THRESHOLD= option, 293
VSCALE= option, 293
WEIBULL Weibull-option, 294
ZETA= option, 294
UNIVARIATE procedure, CLASS statement, 294
  KEYLEVEL= option, 296
  MISSING option, 295
  NOKEYMOVE option, 296
  ORDER= option, 295
UNIVARIATE procedure, FREQ statement, 296
UNIVARIATE procedure, HISTOGRAM statement, 296
  ALPHA= option, 304, 402, 409
  ANNOTATE= option, 491
  BARLABEL= option, 304
  BARWIDTH= option, 304
  BETA option, 305, 401, 481
  BETA= option, 305, 402
  C= option, 305, 418, 419, 488
  CBARLINE= option, 305
  CFILL= option, 305
  CGRID= option, 306
  CLIPCURVES option, 306
  CLIPREF option, 306
  CONTENTS= option, 306
  CPROP= option, 479
  DELTA= option, 306
  EDFNSAMPLES= option, 306
  EDFSIZEED= option, 306
  ENDPOINTS= option, 306, 473
  EXPONENTIAL option, 307, 403
  FILL option, 308
  FRONTREF option, 308
  GAMMA option, 308, 403, 483
  GRID option, 308
  GUMBEL option, 308, 404
  HANGING option, 308
  HOFFSET= option, 309
  IGAUSS option, 309, 405
  INTERBAR= option, 310
  INTERTILE= option, 479
  K= option, 310, 418, 419
  KERNEL option, 310, 418, 419, 488
  LAMBDA= option, 310, 340
  LGRID= option, 310
  LOGNORMAL option, 310, 406, 483, 489, 501
  LOWER= option, 310
  MAXNBIN= option, 310
  MAXSIGMAS= option, 311
  MIDPERCENTS option, 311, 476
  MIDPOINTS= option, 311, 471, 473
  MU= option, 312, 476
  NENDPOINTS= option, 312
  NMIDPOINTS= option, 312
  NOBARS option, 312
  NOPLOT option, 312
  NOPRINT option, 312
  NORMAL option, 312, 406, 476
  NOTABCONTENTS option, 312
  NROWS= option, 466
  OUTHISTOGRAM= option, 313, 431, 473
  OUTKERNEL= option, 432
  PARETO option, 313, 407
  PERCENTS= option, 313
  PFILL= option, 313
  POWER option, 313, 408
  RAYLEIGH option, 314, 409
  RTINCLUDE option, 314, 473
  SB option, 314, 410
  SCALE= option, 403–405, 481
  SIGMA= option, 314, 402, 409, 476
  SU option, 315, 411
  THETA= option, 315, 402, 409, 481, 489, 501
  THRESHOLD= option, 315, 404
  UPPER= option, 315
  VOFFSET= option, 315
  VSCALE= option, 315
  WBARLINE= option, 315
  WEIBULL option, 315, 413, 483
  WGRID= option, 315
  ZETA= option, 316
UNIVARIATE procedure, ID statement, 316
UNIVARIATE procedure, INSET statement, 316
  CFILL= option, 324
  CFILLH= option, 324
  CFRAVE= option, 324
  CHEADER= option, 324
  CSHADOW= option, 324
  CTEXT= option, 324
  DATA option, 325
  DATA= option, 319
  FONT= option, 325
  FORMAT= option, 325
<table>
<thead>
<tr>
<th>Option</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUTTER= option</td>
<td>325</td>
</tr>
<tr>
<td>HEADER= option</td>
<td>325</td>
</tr>
<tr>
<td>HEIGHT= option</td>
<td>325</td>
</tr>
<tr>
<td>NCOLS= option</td>
<td>325</td>
</tr>
<tr>
<td>NOFRAME option</td>
<td>325</td>
</tr>
<tr>
<td>POSITION= option</td>
<td>325</td>
</tr>
<tr>
<td>REFPPOINT= option</td>
<td>325</td>
</tr>
<tr>
<td>OUT= option</td>
<td>326</td>
</tr>
<tr>
<td>PCTLGROUP= option</td>
<td>329</td>
</tr>
<tr>
<td>PCTLNAME= option</td>
<td>330</td>
</tr>
<tr>
<td>PCTLNDEC= option</td>
<td>330</td>
</tr>
<tr>
<td>PCTLPRE= option</td>
<td>331</td>
</tr>
<tr>
<td>PCTLPPTS= option</td>
<td>331</td>
</tr>
<tr>
<td>UNIVARIATE procedure, OUTPUT statement</td>
<td>326, 367</td>
</tr>
<tr>
<td>UNIVARIATE procedure, plot statements</td>
<td>369</td>
</tr>
<tr>
<td>ALPHADELTA= gamma-option</td>
<td>369</td>
</tr>
<tr>
<td>ALPHAINITIAL= gamma-option</td>
<td>369</td>
</tr>
<tr>
<td>ANNOKEY option</td>
<td>369</td>
</tr>
<tr>
<td>ANNOTATE= option</td>
<td>369</td>
</tr>
<tr>
<td>CAXIS= option</td>
<td>369</td>
</tr>
<tr>
<td>CFALSE= option</td>
<td>369</td>
</tr>
<tr>
<td>CFRAPOPTION= option</td>
<td>369</td>
</tr>
<tr>
<td>CHREF= option</td>
<td>370</td>
</tr>
<tr>
<td>COLOR= option</td>
<td>370</td>
</tr>
<tr>
<td>CONTENTS= option</td>
<td>370</td>
</tr>
<tr>
<td>CPROP= option</td>
<td>370</td>
</tr>
<tr>
<td>CSTATREF= option</td>
<td>370</td>
</tr>
<tr>
<td>CTEXT= option</td>
<td>370</td>
</tr>
<tr>
<td>CTEXTSIDE= option</td>
<td>370</td>
</tr>
<tr>
<td>CTEXTTTOP= option</td>
<td>370</td>
</tr>
<tr>
<td>CVREF= option</td>
<td>371</td>
</tr>
<tr>
<td>DESCRIPTION= option</td>
<td>371</td>
</tr>
<tr>
<td>FITINTERVAL= option</td>
<td>371</td>
</tr>
<tr>
<td>FITMETHOD= option</td>
<td>371</td>
</tr>
<tr>
<td>FITTOLERANCE= option</td>
<td>371</td>
</tr>
<tr>
<td>FONT= option</td>
<td>371</td>
</tr>
<tr>
<td>HAXIS= option</td>
<td>371</td>
</tr>
<tr>
<td>HEIGHT= option</td>
<td>371</td>
</tr>
<tr>
<td>HMINOR= option</td>
<td>371</td>
</tr>
<tr>
<td>HREF= option</td>
<td>372</td>
</tr>
<tr>
<td>HREFLABELS= option</td>
<td>372</td>
</tr>
<tr>
<td>HREFLABPOS= option</td>
<td>372</td>
</tr>
<tr>
<td>INFONT= option</td>
<td>372</td>
</tr>
<tr>
<td>INHEIGHT= option</td>
<td>372</td>
</tr>
<tr>
<td>INTERTILE= option</td>
<td>372</td>
</tr>
<tr>
<td>ITPRINT option</td>
<td>372</td>
</tr>
<tr>
<td>L= option</td>
<td>373</td>
</tr>
<tr>
<td>LHREF= option</td>
<td>373</td>
</tr>
<tr>
<td>LSTATREF= option</td>
<td>373</td>
</tr>
<tr>
<td>LYREF= option</td>
<td>373</td>
</tr>
<tr>
<td>MAXITER= option</td>
<td>373</td>
</tr>
<tr>
<td>NAME= option</td>
<td>373</td>
</tr>
<tr>
<td>NCOLS= option</td>
<td>373</td>
</tr>
<tr>
<td>NOFRAME option</td>
<td>373</td>
</tr>
<tr>
<td>NOHLABEL option</td>
<td>373</td>
</tr>
<tr>
<td>NOVLABEL option</td>
<td>374</td>
</tr>
<tr>
<td>NOVTICK option</td>
<td>374</td>
</tr>
<tr>
<td>NROWS= option</td>
<td>374</td>
</tr>
<tr>
<td>ODSFOOTNOTE2= option</td>
<td>374</td>
</tr>
<tr>
<td>ODSFOOTNOTE= option</td>
<td>374</td>
</tr>
<tr>
<td>ODSFOOTNOTES= option</td>
<td>280, 375</td>
</tr>
<tr>
<td>OLED= option</td>
<td>279</td>
</tr>
<tr>
<td>SHAPE= option</td>
<td>375</td>
</tr>
<tr>
<td>VREF= option</td>
<td>377</td>
</tr>
<tr>
<td>VLENGTH= option</td>
<td>377</td>
</tr>
<tr>
<td>VLENGTHPOS= option</td>
<td>377</td>
</tr>
<tr>
<td>XAXIS= option</td>
<td>376</td>
</tr>
<tr>
<td>XAXISLABEL= option</td>
<td>376</td>
</tr>
<tr>
<td>XMINOR= option</td>
<td>376</td>
</tr>
<tr>
<td>YAXIS= option</td>
<td>377</td>
</tr>
<tr>
<td>YAXISLABEL= option</td>
<td>377</td>
</tr>
<tr>
<td>YMINOR= option</td>
<td>377</td>
</tr>
<tr>
<td>ZETA= option</td>
<td>341</td>
</tr>
<tr>
<td>UNIVARIATE procedure, PROBPLOT statement</td>
<td>345</td>
</tr>
<tr>
<td>UNIVARIATE procedure, GUMPLOT statement</td>
<td>332</td>
</tr>
</tbody>
</table>
Syntax Index

CGRID= option, 351
EXPONENTIAL option, 351, 423
GAMMA option, 351, 423
GRID option, 352
GUMBEL option, 352
Gumbel option, 424
LGRID= option, 352
LOGNORMAL option, 352, 424, 497
MU= option, 352
NADJ= option, 352
NORMAL option, 352, 424
PARETO option, 353
Pareto option, 425
PCTLMINOR option, 353
PCTORDER= option, 353
POWER option, 353
power option, 425
RANKADJ= option, 353
RAYLEIGH option, 353
Rayleigh option, 425
ROTATE option, 353
SIGMA= option, 354, 497
SLOPE= option, 354
SQUARE option, 355, 497
THETA= option, 355
THRESHOLD= option, 355
WEIBULL option, 355, 426
WEIBULL2 option, 426
WEIBULL2 statement, 355
WGRID= option, 356
ZETA= option, 356

UNIVARIATE procedure, PROC UNIVARIATE statement, 274
   ALL option, 275
   ALPHA= option, 275
   ANNOTATE= option, 276, 429
   CIBASIC option, 276, 455
   CIPCTLD option, 276
   CIPCTLNORMAL option, 276
   CIQUANTDF option, 457
   CIQUANTNORMAL option, 276, 457
   DATA= option, 277, 429
   EXCLNPWT option, 277
   FREQ option, 277, 445
   GOUT= option, 277
   IDOUT option, 277
   KEYPLEVEL= option, 469
   LOCCOUNT option, 277, 462
   MODES option, 277, 441
   MU0= option, 277
   NEXTRBS= option, 278, 443
   NEXTRVAL= option, 278, 443
   NOBYPLOT option, 278
   NOPRINT option, 278
   NORMAL option, 278
   NORMALTEST option, 278
   NOTABCONTENT= option, 278
   NOVARCONTENT= option, 278
   OUTTABLE= option, 278, 433
   PCTLDEF= option, 278, 382
   PLOT option, 447
   PLOTS option, 278
   PLOTSIZE= option, 280
   ROBUSTSCALE option, 280, 460
   ROUND= option, 280
   SUMMARYCONTENT= option, 280
   TRIMMED= option, 281, 460
   VARDEF= option, 281
   WINSORIZED= option, 281, 460

UNIVARIATE procedure, QQPLOT statement, 356
   ALPHA= option, 356
   BETA option, 361, 422
   BETA= option, 362
   C= option, 362, 516
   CGRID= option, 362
   EXPONENTIAL option, 362, 423
   GAMMA option, 362, 423
   GRID option, 362, 364, 512
   GUMBEL option, 363
   Gumbel option, 424
   LABEL= option, 364
   LGRID= option, 363, 364
   LOGNORMAL option, 363, 424
   MU= option, 363, 505
   NADJ= option, 363, 420
   NORMAL option, 363, 424
   PARETO option, 363
   Pareto option, 425
   PCTLAXIS option, 364, 428, 512
   PCTLMINOR option, 364
   PCTLSCALE option, 364, 428
   POWER option, 364
   power option, 425
   RANKADJ= option, 365, 420
   RAYLEIGH option, 364
   Rayleigh option, 425
   ROTATE option, 365
   SIGMA= option, 365, 505, 508
   SLOPE= option, 365
   SLOPE= option, 365
   SQUARE option, 365
   SQUARE option, 366, 505
   THETA= option, 366
   THRESHOLD= option, 366
   WEIBULL option, 366, 426, 516
   WEIBULL2 option, 426
   WEIBULL2 statement, 366
   WGRID= option, 367
   ZETA= option, 376, 508

UNIVARIATE procedure, VAR statement, 367
UNIVARIATE procedure, WEIGHT statement, 367
UPPER= option
   HISTOGRAM statement (UNIVARIATE), 315
URC option
   OUTPUT statement (FREQ), 104
VAR statement
   CORR procedure, 18
   UNIVARIATE procedure, 367
VAR= option (BINOMIAL)
   TABLES statement (FREQ), 112
VAR= option (RISKDIFF)
   TABLES statement (FREQ), 141
VARDEF= option
   PROC CORR statement, 16
   PROC UNIVARIATE statement, 281
VAXISLABEL= option
   plot statements (UNIVARIATE), 376
VMINOR= option
   plot statements (UNIVARIATE), 377
VOFFSET= option
   HISTOGRAM statement (UNIVARIATE), 315
VREF= option
   plot statements (UNIVARIATE), 377
VREFLABELS= option
   plot statements (UNIVARIATE), 377
VREFLABPOS= option
   plot statements (UNIVARIATE), 377
VSCALE= option
   HISTOGRAM statement (UNIVARIATE), 315
W= option
   plot statements (UNIVARIATE), 377
WARN= option (CHISQ)
   TABLES statement (FREQ), 115
WAXIS= option
   plot statements (UNIVARIATE), 377
WBARLINE= option
   HISTOGRAM statement (UNIVARIATE), 315
WEIBULL option
   HISTOGRAM statement (UNIVARIATE), 315, 413, 483
   PROBPLOT statement (UNIVARIATE), 355, 426
   QQPLOT statement (UNIVARIATE), 366, 426, 516
WEIBULL2 option
   PROBPLOT statement (UNIVARIATE), 355, 426
   QQPLOT statement (UNIVARIATE), 366, 426
WEIGHT statement
   CORR procedure, 18
   FREQ procedure, 145
   UNIVARIATE procedure, 367
WGRID= option
   HISTOGRAM statement (UNIVARIATE), 315
PROBPLOT statement (UNIVARIATE), 356
QQPLOT statement (UNIVARIATE), 367
WINSORIZED= option
   PROC UNIVARIATE statement, 281, 460
WITH statement
   CORR procedure, 19
WTKAP option
   EXACT statement (FREQ), 91
   OUTPUT statement (FREQ), 104
   TEST statement (FREQ), 145
ZELEN option
   EXACT statement (FREQ), 88
   OUTPUT statement (FREQ), 98
ZEROS option
   WEIGHT statement (FREQ), 146
ZETA= option
   HISTOGRAM statement (UNIVARIATE), 316
   PROBPLOT statement (UNIVARIATE), 356
   QQPLOT statement (UNIVARIATE), 367, 508