SAS® Tasks in SAS® Enterprise Guide® 7.1 and SAS® Add-In 8 for Microsoft Office

SAS® Documentation
July 12, 2018

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS® Tasks in SAS® Enterprise Guide® 7.1 and SAS® Add-In 8 for Microsoft Office*. Cary, NC: SAS Institute Inc.

SAS® Tasks in SAS® Enterprise Guide® 7.1 and SAS® Add-In 8 for Microsoft Office

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

July 2018

8.0-P1:egamotasks

Contents

1

Working with Tasks

Built-In Tasks and SAS Studio Tasks

SAS Enterprise Guide and SAS Add-In for Microsoft Office are shipped with many built-in SAS tasks. You can access these tasks from the **Tasks** pane. This list of tasks can also include any custom tasks and any SAS Studio tasks.

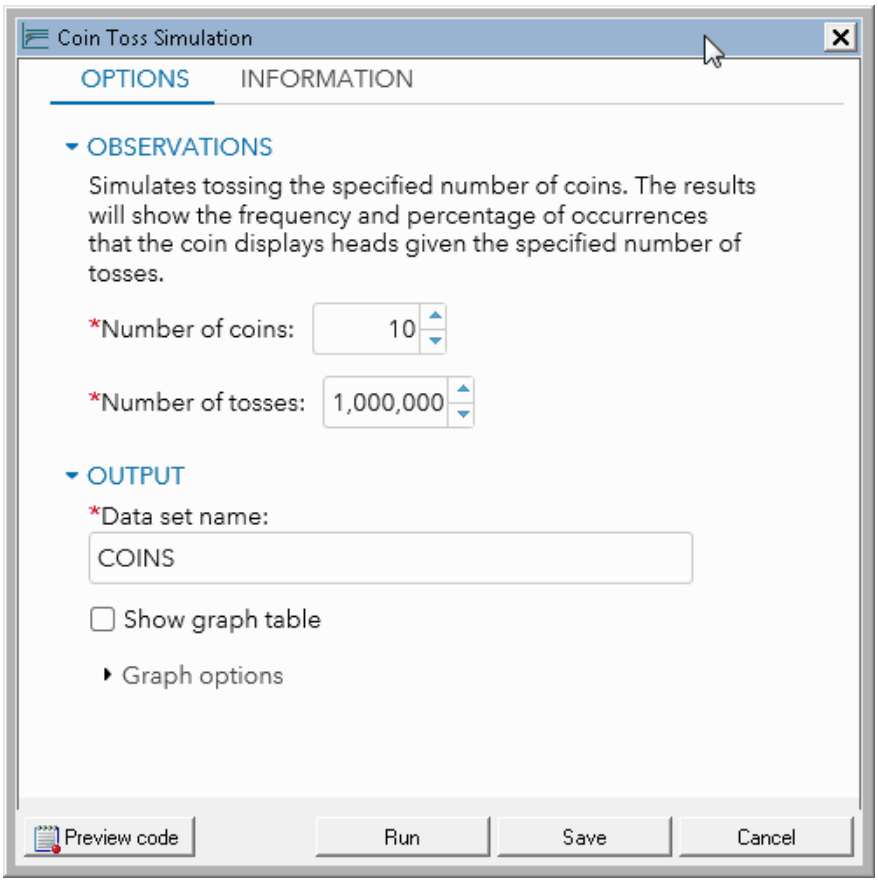
To access SAS Studio tasks in SAS Enterprise Guide or SAS Add-In for Microsoft Office:

1. Verify that you have SAS Studio installed at your site. If you are running SAS Studio 3.4 or later, you can access either the SAS Studio Enterprise Edition or the SAS Studio Single User Edition. The SAS Studio Single User Edition must be installed on the same machine as SAS Enterprise Guide and SAS Add-In for Microsoft Office.
2. Verify that the default server for the active profile is the SAS Studio server.
3. In the settings for SAS Enterprise Guide or SAS Add-In for Microsoft Office, select the **Display SAS Studio tasks** option.

Now, any SAS Studio tasks are available from the **Tasks** pane and Tasks window in SAS Enterprise Guide and SAS Add-In for Microsoft Office. To see only the **SAS Studio** tasks in the **Tasks** pane and task gallery, select **SAS Studio** as the task type.

The interface for SAS Studio tasks is slightly different from the dialog boxes used for the built-in tasks in SAS Enterprise Guide and SAS Add-In for Microsoft Office. For example, here is the interface for the Coin Toss Simulation task.

2



The availability of certain SAS Studio tasks depends on what you license at your site. For example, many SAS Studio tasks enable you to work in a SAS Viya environment. However, if SAS Viya is not licensed and installed, these tasks are not available.

See Also

- "Built-In Tasks" on page 561
- "SAS Studio Tasks and Associated Procedures and Licenses" in SAS Studio: Task Reference Guide

Adding User Code to a Task

When you preview task code, you have the option to insert your own code into the code that SAS generates for the task. You can insert SAS code procedure options or statements into the generated code with all the features of the Program Editor. You cannot edit or delete the automatically generated code within the Code Preview for Task window; you can insert only new code. To insert your own code in the task window, click **Preview code**.

1 In the Code Preview for Task window, select the **Show custom code insertion points** check box. The following lines are added to the task code in each location where you can insert custom code:

```
/* Start of custom user code (Framework_BeforeTaskCode) */
<insert custom code here>
/* End of custom user code (Framework_BeforeTaskCode) */
```

2 Click the location where you want to insert the code and enter your code.

3 To save your changes, close the Code Preview for Task window.

Assigning Column Properties

In the selection pane of the task window, click **Data** to access these options.

Right-click the variable and select **Properties** to temporarily assign a descriptive label and a format. Any properties that you assign apply only to the current task.

To assign a label, enter the label text in the **Label** box. The label cannot exceed 256 characters. The drop-down list holds the value of the label that is permanently associated with the variable.

Most tasks allow you to assign formats. To assign a format:

1 Click **Change**. The Formats window appears.

2 Select the category of the format in the **Categories** box. You can use a standard SAS format or user-written format that you previously defined in the Create Format task.

3 Select the name of the format in the **Formats** box.

4 Specify any changes to the attributes in the **Attributes** box and click **OK**.

Note: In the Properties window, the **Sorted** property is set to `Yes` if the data set has been sorted by that variable.

Selecting Graphics Output Format

In SAS Add-In for Microsoft Office, you can choose from the following graphics output formats: ActiveX, ActiveX image, GIF, JPEG, PNG, and Java image. In **SAS Enterprise Guide**, you can also choose the Java and SAS EMF formats.

The ActiveX and Java output formats are interactive. You can right-click any graphics that are generated in these output formats and change certain options. The options that you can change are different for the ActiveX and the Java output formats. Any options that you change are reflected in the output only, not in the task window selections that you made to create the graph.

The ActiveX image, GIF, JPEG, PNG, Java image, and SAS EMF output formats are noninteractive. You cannot change the appearance of the graph after you generate it.

Working with Task Templates

About Task Templates

When you run a task, such as the Bar Chart task, you might want to frequently use the same settings. Instead of opening the task multiple times to specify these settings for different input data sources, task templates enable you to save your settings for a specific task to a template. You can then run that template with any input data source.

Note: When working with task templates, remember these limitations:

- Values or settings that depend on the data source are not saved in the task template. For example, when you assign a variable to a specific role, the variable assignment is not saved in the task template. If an option is dependent on the values of the data source, then the option setting is not saved in the task template. For example, in the Line Plot task, symbol definitions that depend on the data values are not saved in the task template.
- You cannot create a task template from the following tasks: Append Table, Compare Data, Delete Data Sets and Formats, Download Data Files to PC, Forecast Studio Create Project, Forecast Studio Open Project, Forecast Studio Override Project, List Report, Model Scoring, and Summary Tables.
- In SAS Enterprise Guide, if you use a prompt in the task and then create a template of the task, the prompt is not saved as part of the task template.

Note: In SAS Enterprise Guide you can also create a template from the Query Builder.

In SAS Enterprise Guide, you can manage your templates by using the Task Template Manager.

Create a Task Template

1. Open the task and specify the options for the template.
2. Click the down arrow next to **Run** and select **Create Template** from the pop-up menu. The Create Task Template dialog box appears.
3. Specify a name and description for the template.
4. Specify whether to create a new task template in a group or to replace an existing template.
5. Click **Create**.

Edit a Task Template

You can make changes to templates that you have created. If you are making changes to a template that is based on a query, you must first verify that the data references are still valid.

1. Open the task template that you want to edit. The task window opens with the selections that were specified in the template.
2. Specify the options in the task that you want to save and select **Run** ⇨ **Create Template**. The **Create Task Template** dialog box appears.
3. Select **Replace an existing task template**, and then select the existing template that you want to replace. Click **Create**. Your changes are saved to the existing task template.

Run a Task Template

1. Open a task template by selecting **Tasks** ⇨ **Task Templates** ⇨ ***name-of-group*** ⇨ ***name-of-template***.
2. (Optional) Select the input data source. By default, the active data source in the project is the input data source for the task. However, for most tasks, you can change the input data source from the **Data** panel.
3. In the **Data** panel of the task window, assign variables from the input data source to task roles.
4. (Optional) Modify any additional options in the task.
5. Click **Run**.

Import a Task Template

1. In the **Tasks** pane, click **Import Template**.
2. To select one or more templates to import, click **Add**.
3. From the **Import to group** drop-down list, select the group where you want to place the imported templates.
4. Click **Import**. The templates now appear in the **Tasks** pane and are available from the **Tasks** menu.

Export a Task Template

1. In the **Tasks** pane, click **Export Template**.
2. Select the templates to export and add them to the **Selected task templates** box.
3. To select the location for these templates, click **Export to**.
4. Click **Export**. A copy of the templates is saved to the location that you selected.

2

Append Table Task

About the Append Table Task

The Append Table task concatenates two or more input tables to create a single output data set or SQL view. You can concatenate a maximum of 256 tables. The resulting data set contains every row and every variable that is in any of the input tables. A variable that is included in some but not all of the input tables will have a missing value in each row of the results that come from an input table that does not contain the variable.

For example, you might have four input data sets that show a product's sales numbers for each quarter in 2002. To create a single output data set that shows all of the sales for 2002, you could use the Append Table task.

Note: The Append Table task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SQL
Required SAS products	Base SAS
Recommended additional SAS products	None

Append Table: Building a List of Input Tables

In the selection pane, click **Tables** to select and sequence the tables to be concatenated.

The **Tables to append** area lists the tables that will be concatenated, in the order in which they will be concatenated. Use this area to build the desired list of tables.

1. Click **Add Table**. The Open Data dialog box opens. Select the location of the table.
2. In the dialog box that opens, select the name of the table to add. To select more than one table, hold down the CTRL key while selecting the tables that you want to add.

3 Click **Open** or **OK** to add the table to the **Tables to append** area.

Note: You cannot delete the table that you used to open the Append Table task.

Append Table: Saving Results

In the selection pane, click **Results** to create a SAS data set or an SQL view from your data.

In the **Result set format** area, you can choose whether to save your appended data results as a SAS data set or as an SQL view.

- To save the results as a SAS data set, click **Data table**. This is the default.
- To save the results as an SQL view, click **Data view**.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

3

Area Plot

About the Area Plot Task

The Area Plot task creates area, spline, step, or overlay plots that show the mathematical relationship between two variables. The area under the curve is filled with a pattern or color combination to make a visual impact.

For more information, see "Selecting Graphics Output Format" on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GPLOT
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Area Plot: Selecting a Plot Type

In the selection pane, click **Area Plot** to access these options.

Role Name	Description
Area Line Plot	creates a line plot by using the values of two variables. The area below the plot line is shaded.
Area Spline Plot	creates a line plot by using the spline routine. The area below the spline plot line is shaded.
Area Step Plot	creates a line plot by using the step function. The area below the step plot line is shaded.
Multiple vertical column area plots using overlays	creates a separate plot for each Y variable versus the same X variable. Because of the overlay option, all plots appear on the same graph. The area below each plot is shaded with a different color.
Multiple area plots by group column	creates a separate plot for each unique value of the variable that is assigned to the **Group** role. All plots appear on the same graph. The area below each plot is shaded with a different color.

Area Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart.
Vertical (Right)	The column that you assign to this role is the vertical or Y axis variable for the right side of the graph. This variable is plotted against the X axis variable so that another plot is produced on the graph.
Group	The values of the column that you assign to this role determine how many plot lines are drawn on the same graph.

Role Name	Description
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Select the **Produce same axis scaling for all graphs** check box if you want the scale for each axis to be the same for all graphs that are produced. By default, the range of values for each axis is based on the minimum and maximum values in the data that is being plotted, and therefore it could vary from graph to graph. If you select this option, then the range for each axis is the same for all the graphs.

Area Plot: Setting Appearance Options

Customizing the Area

In the selection pane under the Appearance heading, click **Area** to access these options.

Select an area fill style:

- **Empty**
- **Solid**
- **Line**

If you select **Line**, select the pattern for the fill style and line density.

Note: Not all the area fill styles are supported by all graph output formats. Selecting **Empty** or **Line** as the area fill style has no effect if you are using the ActiveX, Java, ActiveX Image, or Java Image format. These graph output formats display the area plots as if the area fill style is solid.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.

Role Name	Description
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. Note: This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Note: If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.

Role Name	Description
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.

Role Name	Description
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data

- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

4

ARIMA Modeling and Forecasting Task

About the ARIMA Modeling and Forecasting Task

ARIMA forecast modeling analyzes and forecasts equally spaced univariate time series data, transfer function data, and intervention data by using the autoregressive integrated moving average (ARIMA) or autoregressive moving average (ARMA) model. An ARIMA model predicts a value in a response time series as a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series.

There are three basic stages in performing an ARIMA analysis:

- Stage 1: Identifying the time series
- Stage 2: Estimating the time series
- Stage 3: Forecasting future values of the time series.

You might use the ARIMA Modeling and Forecasting task to perform these analyses:

- determine the impact of seasonal factors. For example, you can model a time series with weather variables (temperature, rainfall, humidity, etc.) as inputs.
- forecast the demand for a product based on historical patterns.

Note: The ARIMA Modeling and Forecasting task is organized so that each of the three stages is on a separate page. When you click **Run**, only the stage that corresponds to the selected page and the preceding stages are performed. Thus, if you go to the **Stage 3: Forecasting** page and make changes, and then go to the **Stage 1: Identification** page and click **Run**, only the Identification stage is performed.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	ARIMA
Required SAS products	Base SAS, SAS/ETS
Recommended additional SAS products	SAS/GRAPH

ARIMA Modeling and Forecasting: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Time series variable	specifies the variable to be forecasted. You must assign exactly one variable to this role.
Time ID variable	specifies the variable that is used to assign dates to the rows. You must assign exactly one variable to this role. The first variable that contains a SAS date, time, or datetime format is automatically assigned to this role. You can also create a time ID variable.
Group forecasts by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.

ARIMA Modeling and Forecasting: Setting Identification Options

Differencing the Response Series

In the selection pane under the Stage 1: Identification heading, click **Differencing** to access these options.

To perform differencing on the response series, select the **Difference the response series** check box. When you select this check box, first differences are performed by default. To change the degree of differencing, type the new values in the **Differencing lags** box. For example, specify 1,1 to perform differencing twice at lag one.

Specifying Stationary Tests

In the selection pane under the Stage 1: Identification heading, click **Stationarity** tests to access these options.

To specify a stationarity test, select the option button for the desired test. By default, no tests are performed. If you select a stationarity test, then you can specify the autoregressive orders for the test. By default, autoregressive orders of 0, 1, and 2 are specified. In the Autoregressive orders box, enter the desired autoregressive orders, separated by commas, or type a range of orders, separated by an ellipsis. For example, enter `1...3` to specify autoregressive orders of 1, 2, and 3.

If you select **Perform augmented Dickey-Fuller tests**, you can specify a seasonal lag to perform a seasonal Dickey-Fuller test. Specify an integer value in the **Seasonal lag** box. The default for the seasonal lag is `1`, which is a nonseasonal test.

Generating Plots and Saving the Results

In the selection pane under the Stage 1: Identification heading, click **Plots** and results to access these options.

Option Name	Description
Actual values plot	displays a plot of the input time series in the results. You can also specify the number of lags to consider in computing the autocorrelations and cross-correlations. The default number of lags is `24` or one-fourth the number of observations, whichever is less
Save autocorrelations and cross-covariances	saves an output table of autocorrelations and cross-covariances. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress displayed identification output	suppresses the identification output from the results.

ARIMA Modeling and Forecasting: Setting Estimation Options

Enabling the Estimation Steps

In the selection pane under the Stage 2: Estimation heading, click **Enable estimation steps** access these options.

To activate the options for the estimation steps, select the **Perform estimation steps** check box.

Note: The ARIMA Modeling and Forecasting task is organized into three stages. The options for Stage 1: Identification are always enabled. You must select the appropriate check boxes to activate the options for Stage 2: Estimation and Stage 3: Forecasting. You must activate the estimation steps before you can activate the forecasting steps.

Defining the ARIMA Model

In the selection pane under the Stage 2: Estimation heading, click **Model definition** to access these options.

You can add both autoregressive and moving average parameters to the model estimation.

To add new parameters to the model estimation:

1. In the **Factors for AR model** or **Factors for MA model** box, enter an integer value, a list of integers separated by commas, or a range of integers to represent the lags for the autoregressive or moving average parameters to be estimated for each multiplicative factor in the list. For example, enter `1...3` , `1 TO 3` , or `1,2,3` to specify (1, 2, 3); enter `3` to specify (3).
2. Click **Add** or press Enter to add the new parameter to the list.
3. Repeat this process to add multiplicative terms. For example, enter `1`, click **Add,** enter `12`, and click **Add** to specify (1)(12).

After a parameter is created, you can edit or delete it from the list in the **Autoregressive** or **Moving Average** area.

Selecting the Estimation Method and Options

In the selection pane under the Stage 2: Estimation heading, select **Model options** to access these options.

Option Name	Description
Estimation method	Select an estimation method from the drop-down list. These estimation methods are available: ■ **Maximum likelihood method** produces maximum likelihood estimates. The likelihood function is maximized via nonlinear least squares by using Marquardt's method. ■ **Unconditional least-squares method** produces unconditional least squares estimates. The ULS method is also referred to as the exact least squares (ELS) method. ■ **Conditional least-squares method** produces conditional least squares estimates. These estimates are conditional on the assumption that the past unobserved errors are equal to 0. This method is the default.
Maximum number of iterations	specifies the maximum number of iterations for the estimation stage. Specify an integer value in the **Maximum number of iterations** box. The default value is `50`.
Drop mean term from model	Excludes the mean term from the model.

Saving the Results

In the selection pane under the Stage 2: Estimation heading, select **Results** to access these options.

Option Name	Description
Saving output data	You can save some or all of your results in an output data set. Select the appropriate check boxes to save your results. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress displayed estimation output	Suppresses the output from the estimation stage.

ARIMA Modeling and Forecasting: Setting Forecasting Options

Enabling the Forecasting Steps

In the selection pane under the Stage 3: Forecasting heading, click **Enable forecasting steps** access these options.

To activate the options for the estimation steps, select the **Perform forecasting steps** check box.

Note: The ARIMA Modeling and Forecasting task is organized into three stages. The options for Stage 1: Identification are always enabled. You must select the appropriate check boxes to activate the options for Stage 2: Estimation and Stage 3: Forecasting. You must activate the estimation steps before you can activate the forecasting steps.

Specifying the Forecast Interval

In the selection pane under the Stage 3: Forecasting heading, click **Options** to access these options.

Option Name	Description
Time interval between observations	Select a time interval from the drop-down list to represent the time interval between observations in the input data table. The interval that you choose must be appropriate for the input data, or else the analysis will fail or produce inaccurate results. If you select **Number of units** as the time interval, then you can also specify the time units per interval. The default value is `1`. For any time interval other than **Number of units**, the **Time units per interval** box is unavailable.
Number of intervals to forecast	Specify the number of time intervals for which you want forecasts to be produced. The default value is `24`. For example, if you specify a time interval between observations of month, and if you specify 6 time intervals to forecast, you would get forecasts for 6 months past the end of your input data. Select a confidence level with the **Confidence level** drop-down list. The default is a 95 percent confidence level.

Generating Plots and Saving the Results

In the selection pane under the Stage 3: Forecasting heading, select **Plots and results** to access these options.

Option Name	Description
Forecasting plot options	Generates these plots: ■ To generate a plot of actual values, one-step ahead predicted values, multi-step forecasts, and confidence limits, select the **Forecasts** check box. ■ To generate plots of the residuals, select the **Residuals** check box. Use the **Limit intervals of actual data shown** and the **Intervals to show** boxes to specify the number of intervals of historical data to plot before the forecasted data begins.
Save forecasts	To generate an output table for forecasts, select the **Save forecasts** check box. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress displayed forecasting output	Suppresses the forecasting output from the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

5

Assign Project Library

About the Assign Project Library Wizard

The Assign Project Library wizard enables you to define a library for the project. The wizard generates the LIBNAME statement that you need in order to define the library. This library is then defined whenever you run the task or the process flow that contains the task in the project.

For example, suppose you have some data that you need to analyze. It resides in your local directory and is not in a SAS library that an administrator has defined. You need to define a SAS library so that you can access the data. You can use the Assign Project Library wizard to create this LIBNAME statement.

Note: This task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	Base SAS
Recommended additional SAS products	None

Assign Project Library: Specify a Name and Server

1 .Enter the name of the library in the **Name** box. The name must be all uppercase and can contain a maximum of eight characters. The name must be unique for each server on which the library is created. The name cannot contain any of these characters: \ / : * ? " < > |

2 Select the server from the list of available servers.

3 Click **Next** to select an engine for the library.

Assign Project Library: Specify an Engine

1 Before you can select an engine, you must specify a name and server for the library.

2 Select the engine type. You can choose from these engine types:

Engine Type	Description
File System	Specify these options: ■ To allow SAS to select the engine based on the type of files that are located in the path that you select for the library, select the **Let SAS choose the engine based on the contents of the specified path** check box. **Note:** Do not select this check box if the library path contains mixed file types. ■ Select the engine for the library. If you select **REMOTE-SAS/SHARE** as the engine, you must specify the name-value pair of slibref/SAMPSHR in the specify the configuration options for the library. **Note:** You cannot select an engine if you select the **Let SAS choose the engine based on the contents of the specified path** check box. ■ Enter the physical path where the library resides on the server in the **Path** box. Click **Browse** to select this location. You can use wildcards in the path. <userid> substitutes the Windows user ID <sample> substitutes the Enterprise Client sample directory <serveruser> substitutes the server login user ID <serverpassword> substitutes the server login password <libraryuser> substitutes the library login user ID <librarypassword> substitutes the library login password If credentials are required to access the selected server, you are prompted to enter a user ID and password. **Note:** If you select **BASE**, **V9**, **V8**, **V7**, **V604**, or **V6** as the engine, then you do not have to specify a path. For all other engines, the path is required.

Engine Type	Description
Database System	Specify these options: ■ Select the engine to use when accessing files in the library. ■ Select the external database server for the library. **Note:** The drop-down list displays the database servers that have been defined in the active SAS Metadata Repository. If the server that you want to use is not defined in the metadata, then that server does not appear in the drop-down list. Your administrator can use SAS Management Console to define additional servers. ■ Specify the database schema used to access the data on the server. The schema name that you provide must match the name of a schema that has already been defined on the database server. ■ Enter a user ID and password for the database server.
WebDAV	Specify these options: ■ Select the WebDAV server to use as the library source. **Note:** The drop-down list displays the WebDAV servers that have been defined in the active SAS Metadata Repository. If the server that you want to use is not defined in the metadata, then that server does not appear in the drop-down list. Your administrator can use SAS Management Console to define additional servers. ■ Enter the physical path where the library resides on the server in the **Path** box. You can use these wildcards when specifying the physical path where the library resides on the server or the library options. <userid> substitutes the Windows user ID <sample> substitutes the Enterprise Client sample directory <serveruser> substitutes the server login user ID <serverpassword> substitutes the server login password <libraryuser> substitutes the library login user ID <librarypassword> substitutes the library login password ■ Enter a user ID and password for the WebDAV server.

3 Click **Next** to specify the configuration options for the library.

Assign Project Library: Specify Configuration Options

1 Before you can specify the configuration options, you must select an engine for the library.

2 Specify the library options that have name-value pairs. You can specify only options that have associated values in the table. An example of an option with a name-value pair is ACCESS=READONLY. This option designates the library as read-only so that you cannot write output data to that library. For a list of the available options, see the documentation for your operating environment.

   Note: If you select **REMOTE-SAS/SHARE** as the engine, you must specify the name-value pair of slibref/SAMPSHR in the configuration options for the library.

   To enter an option name or value:

   a In the table, click on a blank cell. When you specify an option name and value, a blank row automatically appears in the table.

   b Enter the name of the option in the **Name** column. You can specify any valid LIBNAME statement option, although you should avoid using prompting options (such as DBPROMPT=YES).

   c .Enter the value that corresponds to the option in the Value column. A blank value is not allowed.

   Note: If the library will connect to a Sybase database, then you must include READLOCK_TYPE=PAGE as an option.

3 Enter any options that do not require values in the **Additional Options** box.

4 Click **Next** to review the summary information.

Assign Project Library: Review the Summary Information

Before creating the new library, you can review the information that you specified.

1 Before you review the summary information, you must specify the configuration options for the library.

2 .Review the information in the text box. To change a value, click **Back**.

3 To test the new library on the server that you selected, click **Test Library**. If SAS Enterprise Guide connects to the server and assigns the library, then **OK** appears as the status of that server. To view the log of the test, click **Show Log**.

4 Click **Finish** to create the new library.

6

Bar Chart

About the Bar Chart Task

The Bar Chart task creates vertical, horizontal, or three-dimensional bar charts that compare numeric values or statistics between different values of a chart variable. Bar charts show the relative magnitude of data by displaying bars of varying height. Each bar represents a category of data.

You might use a bar chart to compare the total amount of sales at each location of a store. In this type of chart, each bar represents the total sales for each site.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GCHART
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Bar Chart: Selecting a Chart Type

In the selection pane, click **Bar Chart** to access these options.

Option Name	Description
Simple	displays the magnitude of data with bars, each of which represents a category of data. The length (or height) of the bars represents the value of the chart statistic for the corresponding midpoint.
Grouped	produces a separate group of bars for each unique value of the variable that is assigned to the **Group bars by** role.
Stacked	creates a separate segment within each bar for every unique value of the variable that is assigned to the **Stack** role. Each segment within a bar is represented by a different color.
Grouped/Stacked	combines the grouped and stacked bar charts. This option produces a separate group of bars for each unique value of the variable that is assigned to the **Group bars by** role and creates a separate segment within each bar for every unique value of the variable that is assigned to the **Stack** role.
Colored	creates a simple bar chart in which each bar is represented by a different color.
Grouped Colored	produces a separate group of bars for each unique value of the variable that is assigned to the **Group bars by** role. Each bar within a group is a different color.
Grouped Colored Groups	produces a separate group of bars for each unique value of the variable that is assigned to the **Group bars by** role. The bars within a group are the same color.
Interleave	produces a separate group of bars for each unique value of the variable that is assigned to the **Group bars by** role. The bars within each group are attached, and each bar is represented by a different color.
3D Grouped	produces a separate group of three-dimensional bars for each unique value of the variable that is assigned to the **Group bars by** role.
3D Grouped/Stacked	combines the grouped and stacked bar charts. This option produces a separate group of three-dimensional bars for each unique value of the variable that is assigned to the **Group bars by** role and creates a separate segment within each bar for every unique value of the variable that is assigned to the **Stack** role.

Option Name	Description
3D Grouped Colored	produces a separate group of three-dimensional bars for each unique value of the variable that is assigned to the **Group bars by** role. Each bar within a group is a different color.
3D Grouped Colored Groups	produces a separate group of three-dimensional bars for each unique value of the variable that is assigned to the **Group bars by** role. The bars within a group are the same color.
Multiple Measures Groups	enables you to assign multiple analysis variables to the Sum of role. The resulting chart contains a group of bars for each unique value of the variable that is assigned to the **Column to chart** role. The bars within a group are the same color.

Bar Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Note: The available roles depend on the chart type that you selected.

Role Name	Description
Column to chart	The values of the column that you assign to this role define the categories of data to chart. You must assign one variable to this role.
Group bars by	The values of the column that you assign to this role determine the groups into which the bars are divided. You must assign one variable to this role.
Stack	The values of the column that you assign to this role determine the number of segments in each bar.
Sum of	The column that you assign to this role determines the lengths of the bars. To choose the particular statistic that is used to determine bar length, select **Advanced** in the selection pane. If you do not assign a column to this role, then the frequency of each value of the **Column to chart** column determines the lengths of the bars or the sizes of the sections.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value.

Bar Chart: Setting Appearance Options

Specifying the Bar Appearance

In the selection pane under the Appearance heading, click **Bars** to access these options.

Role Name	Description
Specifying bar colors	specifies the color of the bars. You can choose from these options: ■ **Default color scheme** uses the default color scheme for the bars. The default color scheme is determined by the current style. You can specify the style by using the Style Manager or the Properties dialog box for the result. In SAS Enterprise Guide, you can also specify the style for each type of result from the Options dialog box. In the SAS Add-In for Microsoft Office, you can apply styles to the SAS Report and HTML formats by using the **SAS Add-In Options** dialog box. ■ **Existing colors (Specified in a previous version)** enables you to use colors that were specified in a previous version of SAS Enterprise Guide or the **SAS Add-In for Microsoft Office**. This option is available only for SAS content that you have migrated from a previous release. ■ **Custom colors** enables you to specify the custom colors to use for the bars. You might want to specify custom colors for bar charts, such as grouped, stacked, and interleaved, that need to use more than one color. Using the drop-down lists, you can specify up to 12 colors for the bar chart. If your chart needs to use more than 12 colors, then the additional colors are selected from the current default color scheme. Click **Reset** to reset the custom colors back to the default. **Note:** Depending on the type of bar chart, not all of the custom colors might be used.
Determining the number of bars	specifies the number of bars in the chart. By default, the number of bars is determined automatically. For numeric columns, the values are divided into ranges, with one bar for each range. If the numeric column represents discrete values, you might want to use the **One bar for each unique data value** option. **Note:** You cannot specify the number of bars for multiple measures bar charts. Select the **Specify number of bars** check box to specify how many bars appear in the chart.

Specifying the Layout

In the selection pane under the Appearance heading, click **Layout** to access these options.

By default, most bar charts are two-dimensional, and the **2D** check box is selected. To create a three-dimensional chart, clear the **2D** check box and select a bar shape from the **Shape** drop-down list. Note

Note: The **2D** check box is not available if you selected a three-dimensional bar chart as the chart type.

Use the **Order** drop-down list to specify the sort order for the bars. By default, the sort order is determined automatically. You can choose to arrange the bars in ascending or descending order of length or height.

Select the **Outline color** check box if you want to outline each of the bars. You can select a color for the outline from the drop-down list.

Use the **Bar size** drop-down list to set the bar size. By default, the bar size is determined automatically. However, you can specify the bar width or spacing between the bars as a percentage of the total chart width in the text box.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. **Note:** This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. **Note:** If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.

Role Name	Description
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. Note: If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. Note: This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. Note: The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Setting Advanced Options

In the selection pane under the Appearance heading, click **Advanced** to access these options.

Role Name	Description
Statistic used to calculate bar	specifies how to calculate the length of the bar. The available statistics depend on the graph that you are creating. ■ If you did not assign a variable to the **Sum of** role in the Bar Chart task or **Bar sum of** role in the Bar-Line Chart task, then the frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. ■ If you assigned a variable to the **Sum of** or **Bar sum of** role, or if you selected multiple measures group as the type of bar chart, then the sum and average statistics are available.

Role Name	Description
Statistic used when plotting line	specifies the statistic to use when plotting the line on a bar-line chart. The frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. If you assigned a variable to the **Line sum of** role in the Bar-Line Chart task, then this option is not available from the **Advanced** panel. Instead, this option is available from the **Plots** panel, so you can control the statistic for each line when more than one variable is assigned to the **Line sum of** role.
Display error bars	displays confidence intervals on the bars. From the **Type** drop-down list, select from these options: ■ **Bars** draws error bars as bars half the width of the main bars. ■ **Both** draws error bars as two tick marks that are joined by a line. ■ **Top** draws the error bar as a tick mark for the upper confidence limit that is joined to the top of the bar by a line. Select a color for the error bars, and use the **CLM** box to specify a confidence level. **Note:** You can display error bars only when these are met: ■ you select **Average** or **Percentage** as the statistic that is used to calculate the length of the bars ■ you are not creating a three-dimensional grouped bar chart ■ you have not assigned a variable to the **Stack** role
Accept missing values	specifies to accept missing values as a valid value for the chart column.
Skip zero values	suppresses any values of the chart column for which the bar length is zero.
Calculate percentages and cumulative percentages for each group (G100)	calculates percentages and cumulative percentages separately for each group in grouped bar charts. When you select this option, the individual percentages reflect the contribution of the midpoint to the group; they total 100 percent for each group. **Note:** This option is available only when you assign a variable to the **Group bars by** role for grouped bar charts.
Show statistics next to bar	shows the value of the statistic that is used to calculate the bar. **Note:** This option is available only for horizontal bar charts.
Specify one statistical value to show for bars	shows the value of a statistic that you select from the drop-down list. The statistic can be the same as or different from the statistic that is used to calculate the bar length. If you are creating a two-dimensional bar chart, you can choose to display the value of the statistic inside the bar.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

7

Bar Chart Wizard

About the Bar Chart Wizard

The Bar Chart wizard helps you create vertical or horizontal bar charts, in two or three dimensions, with options for grouped or stacked bars.

Bar Chart Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to assign variables to roles.

Bar Chart Wizard: Assign Variables to Roles

1 Before you can assign variables to roles, you must select the data.

2 Assign variables to the following roles: ◦◦

- **Bars** - The values of the variable that you assign to this role determine the number of different bars. The variable can be either character or numeric. You must assign one variable to this role. You can choose to sort the bars or to specify the variable type:
  - **Sort** – You can specify the sort order of the bars. By default, the task automatically sorts the bars by ascending order of the midpoint value. You can choose to sort the bars by ascending or descending bar height.

- **Variable type** – By default, the number of bars is determined automatically. For numeric columns, the values are divided into ranges, with one bar for each range. If the numeric column represents discrete values, you might want to select the **Discrete** option.

- **Bar height** - The variable or statistic that you assign to this role determines the lengths of the bars.
  - If you assign a variable to this role, you can choose to calculate the sum or average of each bar. By default, the sum is calculated.
  - If you do not assign a variable to this role, you can choose to calculate the frequency, cumulative frequency, percentage, or cumulative percentage of each bar. By default, the frequency is calculated.

Note: If you are creating a horizontal bar chart, then this option is called **Bar length**.

3 (Optional) Assign variables to these roles:

- **Group by** - The values of the variable that you assign to this role determine the groups into which the bars are divided. For a vertical bar chart, the groups of bars are displayed along the X axis. For a horizontal bar chart, the groups of bars are displayed along the Y axis.

  Note: This option is not available if you assign a variable to the **Depth** role.

- **Depth** - The values of the variable that you assign to this role determine the groups into which the bars are divided. The groups of bars are displayed along the Z axis.

  Note: This option is not available if you create a horizontal bar chart or assign a variable to the **Group by** role.

- **Stack by** - The values of the variable that you assign to this role determine the number of segments in each bar in a bar chart.

- **Chart by** - Separate charts are generated for each group. The groups are determined by the values of the variable that you assign to this role.

4 Click **Next** to set the appearance options.

Bar Chart Wizard: Set the Appearance Options

1 Before you can set the chart appearance, you must assign variables to roles.

2 .Select the **3D chart** check box to create a three-dimensional bar chart. By default, the bars are two-dimensional.

3 Specify how the colors are assigned to the bars in the chart from the **Color bars by** drop-down list. The **All bars the same** and **Bar category** options are available by default. If you assign a variable to the **Group by** or **Depth** role, then the **Group variable** or **Depth variable** option is also available.

Note: If you assigned a variable to the **Stack by** role, then this option is not available.

If all the bars are the same color, then select a color from the **Bar color** drop-down list. If each bar category is a different color, then a color scheme is automatically selected by the Bar Chart wizard.

4 If a legend is available, select the **Legend** check box to include it in the chart. Select the position of the legend from the drop-down list.

5 To specify the statistic to display as the label for each bar, select the **Data labels** check box, and then select the statistic to display from the drop-down list. If you did not assign a variable to the **Bar height** role, then the frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. If you assigned a variable to the **Bar height** role, then the sum and average statistics are also available.

6 To edit the labels for each assigned variable, click **Axis Labels**. The Axis labels dialog box appears.

7 To include reference lines on the chart, select the **Use reference lines** check box.

8 To include tick marks on the chart, select the **Tick marks** check box.

9 Click **Next** to specify title and footnotes.

Bar Chart Wizard: Specify the Titles and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1 Before you can specify the titles and footnotes, you must set the appearance options.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

3 Click **Finish** to run the task.

8

Bar-Line Chart

About the Bar-Line Chart Task

The Bar-Line Chart task creates a vertical bar chart with a line plot overlay. The line plot represents the value of a statistic that is calculated for one of the variables in the input data set.

You can use this task to perform these tasks:

- display and compare exact and relative magnitudes
- examine the contribution of each part to the whole
- determine trends and patterns in the data

For example, you can use this task to show the relationship between product sales and costs for each country where the product is sold. The horizontal axis is the midpoint axis and has a bar for each country. The vertical axes are the response axes. Each bar represents the amount of sales for each country. On the vertical right axis, a line plot of the cost of producing and marketing the product in each country is superimposed over the bars. This graph enables you to compare the cost versus sales for each country and across countries.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GBARLINE

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Bar-Line Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Column to chart	The values of the column that you assign to this role define the categories of data to chart. You must assign one variable to this role.
Bar sum of	The column that you assign to this role determines the lengths of the bars. By default, if no column is assigned to this role, then the total number of observations (the frequency statistic) is used to determine the lengths of the bars. If a column is assigned to this role, then the sum of the values for each category determines the lengths of the bars. To choose the particular statistic that is used to determine the bar length, select **Advanced** in the selection pane. For more information, see Setting advanced options.
Line sum of	The columns that you assign to this role specify the variables to use for the line plots. You can assign more than one variable to this role. To choose the statistic that is used to plot each line, select **Plots** in the selection pane. By default, the sum of the values for each category is used to plot each line. For more information, see Setting the line plot options. **Note:** If you do not assign a column to this role, then the total number of observations for each category (the frequency statistic) is used to plot the line. To choose the statistic that is used to plot the line, select Advanced in the selection pane. For more information, see Setting advanced options.
Stack	The values of the column that you assign to this role determine the number of segments in each bar.

Role Name	Description
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value.

Bar-Line Chart: Setting Appearance Options

Specifying the Bar Appearance

In the selection pane under the Appearance heading, click **Bars** to access these options.

Role Name	Description
Specifying bar colors	specifies the color of the bars. You can choose from these options: ■ **Default color scheme** uses the default color scheme for the bars. The default color scheme is determined by the current style. You can specify the style by using the Style Manager or the Properties dialog box for the result. In SAS Enterprise Guide, you can also specify the style for each type of result from the Options dialog box. In the SAS Add-In for Microsoft Office, you can apply styles to the SAS Report and HTML formats by using the **SAS Add-In Options** dialog box. ■ **Existing colors (Specified in a previous version)** enables you to use colors that were specified in a previous version of SAS Enterprise Guide or the **SAS Add-In for Microsoft Office**. This option is available only for SAS content that you have migrated from a previous release. ■ **Custom colors** enables you to specify the custom colors to use for the bars. You might want to specify custom colors for bar charts, such as grouped, stacked, and interleaved, that need to use more than one color. Using the drop-down lists, you can specify up to 12 colors for the bar chart. If your chart needs to use more than 12 colors, then the additional colors are selected from the current default color scheme. Click **Reset** to reset the custom colors back to the default. Note: Depending on the type of bar chart, not all of the custom colors might be used.
Determining the number of bars	specifies the number of bars in the chart. By default, the number of bars is determined automatically. For numeric columns, the values are divided into ranges, with one bar for each range. If the numeric column represents discrete values, you might want to use the **One bar for each unique data value** option. Note: You cannot specify the number of bars for multiple measures bar charts. Select the **Specify number of bars** check box to specify how many bars appear in the chart.

Setting the Bar Options

In the selection pane under the Appearance heading, click **Bar Options** to access these options.

Use the **Order** drop-down list to specify the sort order for the bars. By default, the sort order is determined automatically. You can choose to arrange the bars in ascending or descending order of length or height.

Select the **Outline color** check box if you want to outline each of the bars. You can select a color for the outline from the drop-down list.

Use the **Bar size** drop-down list to set the bar size. By default, the bar size is determined automatically. However, you can specify the bar width or spacing between the bars as a percentage of the total chart width in the text box.

Setting the Line Plot Options

In the selection pane under the Appearance heading, click **Plots** to access these options.

Note: These options are available only if you assign at least one variable to the **Line sum of** role. If you did not assign a variable to the **Line sum of** role, you can use the **Statistic used when plotting line** drop-down list on the **Advanced** panel to select the statistic to use when plotting the line on a bar-line chart.

To control the appearance of the line plot for a specific variable:

1. Select a variable from the list. The variables in the list are the variables that you assigned to the **Line sum of** role.
2. In the **Line** area, specify the line style, the line width, and the color of the line.
3. Select an outline color.
4. In the **Data point marker** area, specify the type of marker, the symbol to use, the height of each marker, and the color for the data point markers.

   Note: Some symbols are not supported by all graphics output formats. If you select a symbol that is not supported by the output format, then a different symbol (one that is supported) is displayed instead.
5. From the **Statistic used when plotting line** drop-down list, select the statistic to use when plotting the line on a bar-line chart. You can choose from these statistics:
   - **Average** is the average value of the variable that is assigned to the **Line sum of** role. This average is calculated for each unique value, which is represented by a point in the line. For example, to display the average sales for each city in the data set, assign the Sales variable to the **Line sum of** role and select **Average** as the statistic to use for plotting the line.
   - **Sum** is the total of the values for the variable that is assigned to the **Line sum of** role. This total is calculated for each unique value, which is represented by a point in the line. For example, if you assign Sales to the **Line sum of** role, and the values of the Sales variable for the city of Denver are 8734, 982, 1504, 3207, 4502, 624, and 918, the sum statistic for Denver is 20,471.

Selecting the Interpolation Method

In the selection pane under the Appearance heading, click **Interpolations** to access these options.

To select an interpolation method:

1. Select a variable from the list. These are the variables that you assigned to the **Line sum of** role.
2. Select an interpolation method from the drop-down list. Depending on the interpolation method that you choose, you might need to specify additional settings. The following interpolation methods are available:
   - **Line** connects data points with straight lines. Points are connected in the order in which they occur in the input data.
   - **Scatter** suppresses any interpolation. A plot of the data points is created.

- **Step** plots the data by using a step function.

3 (Optional) Select the **Apply to all** check box to use the same interpolation method and property settings for all the variables in the list. This option is available only if there is more than one variable in the list.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.
Axis	customizes the horizontal or vertical axis. - For some plots, you can also set options for the Vertical Right axis. - For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. - For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: - For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. - To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. **Note:** This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. - To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. **Note:** If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. - To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.

Role Name	Description
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. Note: If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. Note: This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. Note: The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Setting Advanced Options

In the selection pane under the Appearance heading, click **Advanced** to access these options.

Role Name	Description
Statistic used to calculate bar	specifies how to calculate the length of the bar. The available statistics depend on the graph that you are creating. ■ If you did not assign a variable to the **Sum of** role in the Bar Chart task or **Bar sum of** role in the Bar-Line Chart task, then the frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. ■ If you assigned a variable to the **Sum of** or **Bar sum of** role, or if you selected multiple measures group as the type of bar chart, then the sum and average statistics are available.

Role Name	Description
Statistic used when plotting line	specifies the statistic to use when plotting the line on a bar-line chart. The frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. If you assigned a variable to the **Line sum of** role in the Bar-Line Chart task, then this option is not available from the **Advanced** panel. Instead, this option is available from the **Plots** panel, so you can control the statistic for each line when more than one variable is assigned to the **Line sum of** role.
Display error bars	displays confidence intervals on the bars. From the **Type** drop-down list, select from these options: ■ **Bars** draws error bars as bars half the width of the main bars. ■ **Both** draws error bars as two tick marks that are joined by a line. ■ **Top** draws the error bar as a tick mark for the upper confidence limit that is joined to the top of the bar by a line. Select a color for the error bars, and use the **CLM** box to specify a confidence level. **Note:** You can display error bars only when these are met: ■ you select **Average** or **Percentage** as the statistic that is used to calculate the length of the bars ■ you are not creating a three-dimensional grouped bar chart ■ you have not assigned a variable to the **Stack** role
Accept missing values	specifies to accept missing values as a valid value for the chart column.
Skip zero values	suppresses any values of the chart column for which the bar length is zero.
Calculate percentages and cumulative percentages for each group (G100)	calculates percentages and cumulative percentages separately for each group in grouped bar charts. When you select this option, the individual percentages reflect the contribution of the midpoint to the group; they total 100 percent for each group. **Note:** This option is available only when you assign a variable to the **Group bars by** role for grouped bar charts.
Show statistics next to bar	shows the value of the statistic that is used to calculate the bar. **Note:** This option is available only for horizontal bar charts.
Specify one statistical value to show for bars	shows the value of a statistic that you select from the drop-down list. The statistic can be the same as or different from the statistic that is used to calculate the bar length. If you are creating a two-dimensional bar chart, you can choose to display the value of the statistic inside the bar.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

9

Basic Forecasting

About the Basic Forecasting Task

The Basic Forecasting task provides a quick and automatic way to generate forecasts for many time series in one step. This task uses extrapolative forecasting methods where the forecasts for a series are functions only of time and past values of the series, not of other variables.

You might want to use this task to forecast sales of a product. For example, suppose that you saved the monthly data on the sales of some product in a data set, and you want to forecast sales for the next 10 months. This task can compute the dates for the forecast values and the confidence limits for the forecasted sales. You can save this output in a SAS data set. You can also save parameters of the forecasting models that are used, as well as statistics that measure how well the forecasting models fit the data, to an output SAS data set. Finally, you can plot the forecasts, confidence limits, residuals, and actual values.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	FORECAST
Required SAS products	Base SAS, SAS/ETS
Recommended additional SAS products	SAS/GRAPH

Basic Forecasting: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Forecast variable	specifies the variables to be forecasted. You must assign exactly one variable to this role.
Time ID variable	specifies the variable that is used to assign dates to the rows. You must assign exactly one variable to this role. The first variable that contains a SAS date, time, or datetime format is automatically assigned to this role. You can also create a time ID variable.
Group forecasts by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already assigned to another role.

Basic Forecasting: Setting Options for Forecasting

Selecting a Forecasting Method and Intervals

In the selection pane, click **Forecast Options** to access these options.

Option Name	Description
Forecasting Method	To specify a forecasting method, select the desired method from the drop-down list. Note ■ **Stepwise autoregressive** (for autoregressive processes) fits a time trend model to the series and takes the difference between each value and the estimated trend. Then, the remaining variation is fit by using an autoregressive model. This method is the default. ■ **Exponential smoothing** (for moving average processes) fits a trend model such that the most recent data are weighted more heavily than data in the early part of the series. The weight of an observation is a geometric (exponential) function of the number of periods that the observation extends into the past relative to the current period. ■ **Winters (Multiplicative) Method** (for multiplicative seasonal processes) uses updating equations similar to exponential smoothing to fit parameters for the model. In this model, you have two trend parameters and a function that selects the seasonal parameter corresponding to the time. ■ **Winters Additive Method** (for additive seasonal processes) is similar to the Winters (multiplicative) method except that the seasonal parameters are added to the trend rather than multiplied with the trend. **Note:** Selecting a forecasting method might enable or disable other controls in this window, as appropriate.
Number of intervals to forecast	Specify the number of time intervals for which you want forecasts to be produced in the **Number of intervals to forecast** box. The default value is `12`. For example, if you specify a time interval between observations of month, and specify 6 lead time intervals to forecast, you would get forecasts for 6 months past the end of your input data.

Selecting Input Data Options

In the selection pane, click **Forecast Options** to access these options.

Option Name	Description
Time interval between observations	specifies the time interval between rows in the input data table. The default time interval is **Number of units**. If you select **Number of units** as the time interval, you can also select the time units per interval. For any other time interval, **Time units per interval** is unavailable. The interval that you choose must be appropriate for the input data, or else the analysis will fail or produce inaccurate results.
Seasonal cycle length	specifies a seasonal cycle length and the intervals per seasonal cycle for a seasonal forecasting method (Winters multiplicative method or Winters additive method). The available seasonal cycle lengths depend on what time interval you select in the box above. The default seasonal cycle length is **Number of intervals**.

Setting Forecasting Method Options

In the selection pane, click **Forecast Options** to access these options.

Plot Name	Description
Degree of time trend model	To specify the degree of the time trend model, select a value from the drop-down list. The available models are **Constant**, **Linear**, and **Quadratic**. By default, all of the forecasting methods use a linear time trend model, except the exponential smoothing method, which uses a quadratic model.
Confidence level	Select a confidence level from the drop-down list. The default is a 95% confidence level.
Component smoothing weights	You can specify smoothing weights for the exponential smoothing, Winters multiplicative, and Winters additive methods. For the exponential smoothing method, you can specify a smoothing weight only for the constant term. For the Winters multiplicative and Winters additive methods, you can specify a smoothing weight for the constant component, the linear and quadratic component, and the seasonal component. The linear and quadratic component becomes available only if you specify a constant smoothing weight. Likewise, you can specify a seasonal smoothing weight only after specifying one for the linear and quadratic component. The default smoothing weights depend on the specified degree of the time trend model and on the other smoothing weights that have been specified. To change a smoothing weight, select the relevant check box and enter a new value in the field.

Basic Forecasting: Generating Plots

In the selection pane, click **Plots** to access these options.

Use the check boxes in the **Types** area to generate plots of the actual values, forecast data, and residuals.

For forecast data plots, you can choose to include this data

- **Actual values and 1-step-ahead predictions** includes the actual data values and the one-step-ahead predictions on the plot. A one-step-ahead prediction refers to a prediction that was made during the previous time period (*t*-1) for one time period in the future.
- **Confidence limits of the forecasts** includes the confidence limit for each forecasted value on the plot.

For all plot types, use the **Limit actual data shown** check box and the **Intervals to show** box to select the number of intervals of historical data to plot before the forecasted data begins.

Note: The **Limit actual data shown** check box is unavailable when you assign a variable to the **Group forecasts by** role.

Basic Forecasting: Setting Results Options

In the selection pane, click **Results** to access these options.

The data set that contains the forecasted values is always saved. To generate a permanently saved table of parameter estimates, select the **Parameter estimates** check box. Select additional statistics to include.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

You can select whether to include forecasts and parameter estimates in the results. You must select the **Parameter estimates** check box in the **Parameter estimates** area before you can select the **Parameter estimates** check box in the **Display output** area.

Note: If you do not select any plots and you do not request that either the forecasts or estimates be shown, then no output will be generated.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

10

Box Chart

About the Box Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The Box Chart task creates a mean chart for the subgroup means. The box plots of the measurements for each subgroup are superimposed onto the mean chart. The box chart provides a detailed analysis of the subgroup distributions and is recommended for large subgroup sample sizes (10 or more).

You might want to use a box chart to find the distribution of the output and to determine whether a process is in statistical control. For example, suppose that a petroleum company uses a turbine to heat water into steam that is pumped into the ground to make oil less viscous and easier to extract. This process occurs 20 times daily, and the amount of power (in kilowatts) that is used to heat the water to the desired temperature is recorded. You can use a box chart to examine the distribution of power output for each day and to determine whether the mean level of the heating process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Control Charts: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Process measurement	specifies the process variables to be analyzed. These variables contain the individual measurements. You must assign at least one variable to this role.
\Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A subgroup identifier is required, and you can assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Subgroup sample size	specifies the subgroup sample sizes as the values of a variable. The subgroup sample size is required, and you assign only one numeric variable to this role. Select the **Use numeric value for subgroup (instead of a variable)** check box when the subgroup sample size is fixed and you need to specify a constant value for the size. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** This role is not required if you select the **Use numeric value for subgroup (instead of a variable)** check box on the **Options** pane.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Box Chart Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame. Note: The **Inside control limits** and **Control limits** options are not available for a box chart if you suppress the control limits. To activate the control limits, clear the **Suppress display of control limits** check box.
Box colors	specifies the outline color and fill color of the boxes.
Suppress display of control limits	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper right corner of the chart.
Vary box width by subgroup size	specifies to vary the box plots in proportion to the subgroup size.
Notch boxes	creates box plots that are notched. The endpoints of the notches are located at the median, plus and minus the quantity 1.58(IQR/sqrt(*n*)), where IQR is the interquartile range and sqrt(*n*) is the square root of the subgroup sample size. The medians (central lines) of two box plots are significantly different at approximately the 0.05 level if the corresponding notches do not overlap.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.

Role Name	Description
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

11

Box Plot

About the Box Plot Task

The Box Plot task creates box plots, high-low charts, or high-low-close charts that display multiple summary statistics for some numeric variable across different values of a chart variable. For example, a box plot can show the range, interquartile range, and mean of a variable. A high-low-close chart can show the high, low, and closing price of a stock on different dates.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GPLOT
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Box Plot: Selecting a Plot Type

In the selection pane, click **Box Plot** to access these options.

Option Name	Description
Box Plot	specifies that the bottom and top edges of the box are located at the sample 25th and 75th percentiles. The center horizontal line is drawn at the 50th percentile (median). Vertical lines, or whiskers, are drawn from the box to the most extreme point within 1.5 interquartile ranges. Values outside this range are marked with a plot symbol.
Hi-Lo Plot	specifies that a solid vertical line connect the low (minimum) and high (maximum) Y values for each X value. A tick mark indicates the mean Y value for each X value. This type of plot is useful for stock market data.
Hi-Lo-Close Plot	specifies that a solid vertical line connect the low (minimum) and high (maximum) Y values for each X value. A tick mark indicates the close value. This type of plot is useful for stock market data.

Box Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart.
Vertical (Right)	The column that you assign to this role is the vertical or Y axis variable for the right side of the graph. This variable is plotted against the X axis variable so that another plot is produced on the graph.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Select the **Produce same axis scaling for all graphs** check box if you want the scale for each axis to be the same for all graphs that are produced. By default, the range of values for each axis is based on the minimum and maximum values in the data that is being plotted, and therefore it could vary from graph to graph. If you select this option, then the range for each axis is the same for all the graphs.

Box Plot: Setting Appearance Options

Setting Box Plot Options

In the selection pane under the Appearance heading, click **Box Plot** to access these options.

Note: These options are available only if you select a box plot.

Role Name	Description
Drawing median lines and the top and bottom of each whisker	You can choose to draw lines between the medians and to draw the tops and bottoms on each whisker.
Specifying colors for the box and outliers	To select your own combination of fill and outline colors, select the **Specify colors** check box, and select your colors from the **Box fill and outliers** and **Outline** drop-down lists. The outline color specifies the color of the box outline and the whisker lines in high-low charts and box plots.
Including values outside the axis range	To specify the values outside of the axis range are included in the interpolation calculations, select the **Include values outside axis range** check box. When this option is selected, values that fall outside of the axis range are included in interpolation calculations but excluded from the plot. By default, this check box is not selected, and values outside the axis range are omitted from the interpolation calculations.

Role Name	Description
Specifying the length of the whiskers	You can select a length percentile for the whiskers. These options are available: ■ **+- 1.5 times the interquartile range** draws the vertical lines, or whiskers, from the box to the most extreme point within 1.5 interquartile ranges. ■ **high/low extremes** draws the whiskers from the extreme high and low values. **Note:** If you select this option, any values that are outside the range of the length percentile for the whiskers (also called outliers) are not marked with plot symbols. If you select any other option for the length percentile, then the outliers are marked with plot symbols. ■ **1st percentile low, 99th high** draws the whiskers from the extreme values in the 1st percentile to the 99th percentile. ■ **5th percentile low, 95th high** draws the whiskers from the extreme values in the 5th percentile to the 95th percentile. ■ **10th percentile low, 90th high** draws the whiskers from the extreme values in the 10th percentile to the 90th percentile. ■ **25th percentile low, 75th high** produces no whiskers because the box extends from the 25th to the 75th percentile.
Customizing the outliers	You can specify the type of font to use, the symbol, and the height.

Setting High–Low Plot Options

In the selection pane under the Appearance heading, click **Hi-Lo** to access these options.

Note: These options are available only if you select a high–low chart or high–low-close chart.

Role Name	Description
Drawing median lines and the top and bottom of each whisker	You can choose to draw lines between the medians and to draw the tops and bottoms on each whisker.
Connecting the high and low values with empty bars	You can choose to draw empty bars instead of lines between the minimum and maximum Y values.
Drawing tick marks at the close value	You can choose to draw a tick mark at the close value rather than the mean value. To use this option, you must have exactly three values of Y (high, low, and close) for each value of X. If there are more or fewer than three Y values for each X value, the tick mark is placed at the mean of the Y values.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. **Note:** This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. **Note:** If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.

Role Name	Description
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task

- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

12

Bubble Plot

About the Bubble Plot Task

The Bubble Plot task creates simple bubble plots by drawing circles of varying proportions at data points that are plotted on the vertical and the horizontal axes. Two of the variables determine the location of the data points. The third variable controls the size of the circles.

You might want to use a bubble plot to compare salary information. For example, suppose that you create a plot in which each bubble represents a category of engineer, such as aerospace, chemical, civil, electrical, and mechanical. The location of a bubble is determined by the average salary for the category that it represents. The size of a bubble represents the number of engineers in the category relative to the total number of engineers in the data.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GPLOT
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Bubble Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart.
Vertical (Right)	The column that you assign to this role is the vertical or Y axis variable for the right side of the graph. This variable is plotted against the X axis variable so that another plot is produced on the graph.
Bubble size	The column that you assign to this role determines the relative sizes of the bubbles on the plot of the Horizontal and Vertical variables. Note: If you have a variable assigned to the **Vertical Right** role but you did not assign a variable to the **Bubble2 size** role, this role also determines the relative sizes of the bubbles on the plot of the Horizontal and Vertical Right variables.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Select the **Produce same axis scaling for all graphs** check box if you want the scale for each axis to be the same for all graphs that are produced. By default, the range of values for each axis is based on the minimum and maximum values in the data that is being plotted, and therefore it could vary from graph to graph. If you select this option, then the range for each axis is the same for all the graphs.
Bubble2 size	The column that you assign to this role determines the relative sizes of the bubbles on the plot of the Horizontal and Vertical Right variables. Note: This option is available only when you assign a variable to the **Vertical Right** role.

Bubble Plot: Setting Appearance Options

Setting Bubble Plot Options

In the selection pane under the Appearance heading, click **Bubble options** to access these options.

Select a bubble color from the drop-down list.

You can choose to label the bubbles. The label is the value of the column that is assigned to the **Bubble size** role for the combination of Horizontal and Vertical variable values. If this variable has a format, the formatted value is used. By default, bubbles are not labeled.

Use the slider to scale all bubbles to a larger or smaller size. You might want to view the output before adjusting this value.

Note: If you assigned a variable to the **Vertical Right** role, then you can also set the Bubble2 options for the bubbles that are for that combination of Horizontal and Vertical Right variables.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.

Role Name	Description
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. Note: This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Note: If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.

Role Name	Description
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.

Role Name	Description
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

13

Canonical Correlation

About the Canonical Correlation Task

Canonical correlation analysis examines the relationship between a linear combination of a set of X variables and a linear combination of a set of Y variables. A series of hypotheses test that each canonical correlation and all smaller canonical correlations are zero in the population. Simple and multiple correlation are special cases of canonical correlation in which one or both sets contain a single variable.

You might use the Canonical Correlation task to determine the degree of correspondence between a set of job characteristics and a set of measures of employee satisfaction.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CANCORR
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Canonical Correlation: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Set 1 variables	performs analyses using the variables that you assign to this role. You must assign at least one variable to this role. Use the **Set 1 Canonical Variates** area to specify a prefix and label for the Set 1 variables.
Set 2 variables	performs analyses using the variables that you assign to this role. You must assign at least one variable to this role. Use the **Set 2 Canonical Variates** area to specify a prefix and label for the Set 2 variables.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analysis by a variable that you have already assigned as **Set 1 variables** or **Set 2 variables**.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Partial variables	specifies the variables to base the canonical analysis on partial correlations. The variables in this role are partialed out of the Set 1 variables and Set 2 variables.
Relative weight	specifies the value of the variable in each row to compute weighted product-moment correlation coefficients. A row is used only if the value of the variable in this role is greater than zero. You can assign a maximum of one variable to this role.

Canonical Correlation: Setting Statistics Options

In the selection pane, click **Statistics** to access these options.

Option Name	Description
Regression analyses to perform	sets the regression analysis to perform. You can choose from these options: ■ **No regression analyses** does not perform any regression analyses. ■ **Set 1 predicts set 2, and set 2 predicts set 1** requests two multiple regression analyses so that **Set 1 variables** are the dependent variables for one analysis and the regressors for the other. ■ **Set 1 predicts set 2** requests multiple regression analyses with **Set 2 variables** as the dependent variables and **Set 1 variables** as the regressors. ■ **Set 2 predicts set 1** requests multiple regression analyses with **Set 1 variables** as the dependent variables and **Set 2 variables** as the regressors.
Regression statistics	calculates one or more of these statistics: ■ **Regression coefficients** produces raw regression coefficients from the regression analyses. ■ **Standardized regression coefficients** produces standardized regression coefficients. ■ **Standard error coefficients** produces standard errors of the regression coefficients. ■ **t Statistic and Prob > |t|** produces *t* statistics for the regression coefficients and probability levels for the *t* statistics. ■ **Squared multiple correlation** produces squared multiple correlations and *F* tests for the regression analyses.

Option Name	Description
Correlation statistics	calculates one or more of these statistics: ■ **Correlations of regression coefficients** produces correlations among the regression coefficient estimates. ■ **Partial correlations** produces partial correlations between regressors and dependent variables, removing from each dependent variable and regressor the effects of all other regressors. ■ **Squared partial correlations** produces squared partial correlations between regressors and dependent variables, removing from each dependent variable and regressor the effects of all other regressors. ■ **Semipartial correlations** produces semipartial correlations between regressors and dependent variables, removing from each regressor the effects of all other regressors. ■ **Squared semipartial correlation** produces squared semipartial correlations between regressors and dependent variables, removing from each regressor the effects of all other regressors.

Canonical Correlation: Generating Plots

In the selection pane, click **Plots** to access these options.

Note: The plot options are not available if the number of canonical variables for which coefficients and redundancy statistics are shown is zero. For more information, see the Number of canonical variables option in Results.

To display plots, select the **Show plots of canonical variables** check box. Use the **Start variable** and **End variable** boxes to enter the variables to plot. Here are some examples:

- To plot only the first canonical variable (V1 versus W1), enter `1` in both boxes.
- To plot the first, second, and third canonical variables (V1 versus W1, V2 versus W2, and V3 versus W3), enter `1` in the first box and `3` in the second box.

Canonical Correlation: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Save the results in an output data set	saves the results to an output data set. You can choose from these options: ■ **Save correlations** creates a data set that contains various statistics, including the canonical correlations and coefficients and the multiple regression statistics that you request. ■ **Save scores** creates a data set that contains all the original data plus the scores for the canonical variables. The number of new variables is twice the value that is specified in the **Number of canonical variables** box. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Show results	displays the output. You can specify these options: ■ **Number of canonical variables** specifies how many canonical variables will include full output in the results. ■ **Include canonical redundancy analysis** performs canonical redundancy analysis, including raw (unstandardized) and standardized variance and cumulative proportion of the variance of each set of variables, explained by their own canonical variables and explained by the opposite canonical variables. ■ **Include scores** displays the output data set that contains all the original data plus the scores on the canonical variables. If you want to create a permanent data set to use later, then select the **Save scores** check box.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

14

c Chart

About the c Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The c Chart task creates c charts for the numbers of nonconformities (defects) in the subgroup samples.

You might want to use a c chart to monitor the number of defects that are found in a new product. For example, suppose that an automobile company wants to monitor the number of paint defects on its new trucks. Twenty trucks of the same model are inspected, and the number of paint defects per truck is recorded. Each point on the c chart represents the number of defects for a given truck. The upper and lower control limits are marked, so that you can see whether any points are above or below these limits. If none of the points exceed the specified control limit, then the painting process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

c Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Number of nonconformities	specifies the variables that contain the number of nonconformities. You must assign at least one variable to this role
\Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A subgroup identifier is required, and you can assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. Note: To prevent the input data set from being sorted, clear the **Sort by variables** check box. Note: You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.

Option Name	Description
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

15

CDF Plots

About the CDF Plots Task

A process capability analysis compares the distribution of output from a process in statistical control to its specification limits in order to determine the consistency with which the specifications are met. A CDF plot is used in process capability analysis to plot the observed cumulative distribution function (CDF) of a variable. The CDF is also referred to as the empirical cumulative distribution function (ECDF) and is an increasing step function.

You might use a CDF plot to determine the percentage of observations that are below the lower specification limit or above the upper specification limit. For example, suppose that a company that produces fiber-optic cord is interested in the breaking strength of the cord. You create a data set that contains 50 breaking strengths, measured in pounds per square inch. The CDF Plot task uses the data to create a CDF plot that displays the percentage of observations in the sample that are below the lower specification limit.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CAPABILITY
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Capability Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the numeric variables to be analyzed. You must assign at least one variable to this role. For all plots except the P-P plot, you can specify an upper limit, a target value, and a lower limit in the **Spec limits for *variable-name*** area for each analysis variable. When you specify these limits, the output automatically includes a table that shows the tests for normality, a table that shows the specification limits, and a table that shows the process capability indices.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight. When you assign a variable to this role, the value of the variable in each observation is used to compute weighted statistics. You can assign only one variable to this role. **Note:** Assigning a variable to the **Relative weight** role is not recommended when you are creating a histogram, probability plot, P-P plot, Q-Q plot, or CDF plot. If you assign a variable to this role, any output should be interpreted with caution.
Classification variables (available for the Histogram task only)	specifies one or two variables to use as classification variables. When you assign a variable to this role, the task creates a comparative histogram that enables you to compare the distribution of an analysis variable across classification levels. **Note:** To create comparative probability plots or comparative Q-Q plots, use the Distribution Analysis task.

CDF Plots: Selecting the Distribution

In the selection pane under the Distributions heading, click **Summary** to access these options.

You can fit one theoretical distribution on a CDF plot. After selecting the distribution, select the distribution name in the selection pane to specify the parameter values and line properties for that distribution.

Note: If you select multiple distributions, a separate CDF plot is created for each distribution.

Option Name	Description
Distributions	Use the check boxes to select one or more of these estimated density curves: ■ **Normal** fits a normal density with a mean (mu) and a standard deviation (sigma). ■ **Lognormal** fits a lognormal density with a scale parameter (zeta), a threshold parameter (theta), and a shape parameter (sigma). ■ **Exponential** fits an exponential density with a scale parameter (sigma) and a threshold parameter (theta). ■ **Weibull** fits a three-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). ■ **Beta** fits a beta density with a scale parameter (sigma), a threshold parameter (theta), and shape parameters (alpha and beta). ■ **Gamma** fits a gamma density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (alpha).
Graphics style	Select the graphics style to use in the output. You can choose between traditional SAS graphics or the ODS statistical graphics. For more information about ODS graphics, see *SAS Output Delivery System: User's Guide*.

CDF Plots: Customizing the Distribution

In the selection pane under the Distributions heading, click the name of the distribution that you want to customize to access these options.

Note: These distribution options are available only if you previously selected the distribution on the **Summary** page or if you select the ***Distribution name*** check box on the page for that distribution.

You can specify the distribution parameters for each analysis variable or all of the analysis variables.By default, the task estimates the parameter values for all of the analysis variables.

To specify the parameters for each analysis variable:

1. Clear the **Apply distribution on all variables** check box.
2. Select the variable's check box in the **Analysis variables** box.
3. Specify the parameter values.

Note: To return to the estimated parameter values, click **Use Estimates**.

Capability Analysis: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Adding an Inset

In the selection pane under the **Appearance** heading, click **Inset** to access these options.

You can enhance a plot by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative histogram, an inset appears in each component histogram.

Select **Include inset** to add summary statistics to the plot. Then select the statistics that you want to include in the inset.

Note: The statistics for calculating the capability indices are available only if you created specification limits for the variable that you assigned to the **Analysis variables** role. You can create specification limits for histograms, CDF plots, probability plots, and Q-Q plots. For more information about calculating capability indices, see the Help for the CAPABILITY procedure.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

The plot appearance options that are available depend on which type of plot the task produces.

Role Name	Description
Background	specifies the background color for the plot.
Bar outline and **Bar fill**	specifies the colors for the bars in a histogram. **Note:** These options are available only for histograms.

Role Name	Description
Reverse the two axes putting the percentiles on the opposite axis	reverses the axes on the plot. If you select this option, then the axes options and reference lines that you selected might no longer work since the orientation of the plot will have changed. **Note:** This option is available only for probability plots and Q-Q plots.
Suppress default legends	prevents the default legend from being displayed with the plot.
Use midpoints	specify midpoints for the histogram. The options that are available from the drop-down list depend on whether you are creating a histogram or a comparative histogram. If you assign a variable to the **Classification variables** role, you are creating a comparative histogram. For comparative histograms, these options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Key** determines the midpoints for the data in the key cell. By default, the levels for the classification variable are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first classification variable. If you specified only one classification variable, then the key cell is the level that occurs first in this order. If you specified two classification variables, then the key cell is the combination of levels of variable 1 and variable 2 that occurs first in this order. Thus, the choice of the key cell determines the uniform horizontal axis that is used for all cells. The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. ■ **Uniform** determines the number of midpoints based on the total sample size. If you do not assign a variable to the **Classification variables** role, you are creating a histogram. For histograms, the following options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Specify number of bins** enables you to specify the number of bins (also called histogram intervals) for the data. By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. **Note:** You can specify midpoints only if you are using the Histograms task.
Specify custom chart size (in pixels)	specifies the image size. For reference, the total screen size is given in the **Width** and **Height** boxes.

Note: Comparative options are available for comparative histograms. You can create a comparative histogram by assigning a variable to the **Classification variables** role in the Histograms task.

Role Name	Description
Rows and **Columns**	If you assign at least one classification variable to an analysis role, then you can use the **Rows per page** and **Columns per page** boxes to change the arrangement of rows and columns in the comparative histogram. If you assign one classification variable, the default arrangement is two rows and one column per page. If you assign two classification variables, the default arrangement is two rows and two columns per page. You can also use the **Frame side** and **Frame top** drop-down lists to specify the colors that are used to fill the frame area of the row labels and the column labels.
Number of bins	By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. You can also specify a standard deviation such that the number of bins that are displayed is based on a range of the standard deviation value above and below the mean of the data.

Setting Specification Limit Options

Note: These options are available for histograms, probability plots, Q-Q plots, and CDF plots.

In the selection pane under the **Appearance** heading, click **Spec Limits** to specify the lower specification limit, the target, and the upper specification limit.

Note: The **Spec Limits** options are available only if you specify one or more limits for at least one analysis variable. To specify a specification limit for an analysis variable, click **Data** in the selection pane, and then select the analysis variable.

Capability Analysis: Specifying the Types of Analyses

In the selection pane, click **Tables** to access these options.

By default, the Capability Analysis report includes the following tables: a basic confidence intervals table, a basic measures table, a tests for location table, and a moments table.

To add a table to the report , select the table in the **Tables** box.

To specify the analysis options, select the table in the **Tables** box. If options are available, a box appears on the right that you use to specify the options.

To suppress all descriptive statistics tables from the report , select the **Suppress descriptive statistics and capability indices tables** check box. Selecting this option does not suppress these tables:

- tables that are created by the distribution that you selected. You can suppress these tables by selecting the **Suppress distribution tables** check box on the panel for that distribution.
- tables that are created by the INTERVALS statement. To suppress these tables, you can specify the NOPRINT option. For more information about the INTERVALS statement, see the Help for the CAPABILITY procedure.

To save your results to an output data set , select the **Save output statistics to a data set** check box. The data set will include univariate statistics and capability indices.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

16

Characterize Data

About the Characterize Data Wizard

The Characterize Data wizard helps you to create a summary report, graphs, and frequency and univariate SAS data sets that describe the main characteristics of the data.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CONTENTS, FREQ, UNIVARIATE
Required SAS products	Base SAS
Recommended additional SAS products	SAS/GRAPH

Characterize Data: Select the Data

1 Click **Add**. The Open dialog box appears.

2 Select the data set that you want to use. To select multiple data sets, press the CTRL key. Click **Open**. The selected data sets appear in the **Data** box.

3 Click **Next** to select the report options.

Characterize Data: Select the Report Options

1 Before you can select one of the report options, you must select the data.

2 Select the output types. By default, a summary report, graphs, and SAS data sets for the frequency data and univariate data are created.

   If you select **SAS Data Sets** as an output type, then SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options dialog box and saves the output data in the first writable library from that list. To change the location, click **Browse**. The resulting data sets are also added to the project.

3 Click **Next** to specify the output limits.

Characterize Data: Specify the Output Limits

1 Before you can specify the output limits, you must select one of the report options.

2 Select the **Limit categorical values** check box to specify the maximum number of unique categorical values to report per variable. By default, `30` values are reported.

3 Click **Finish** to generate the output.

17

Cluster Analysis

About the Cluster Analysis Task

The Cluster Analysis task creates hierarchical clusters of the observations in a SAS data set that contains either coordinate data or distance data. If the data set contains coordinate data, the task computes Euclidean distances before applying the clustering methods. The task can graph the results of the hierarchical clustering to produce a tree diagram (called a dendrogram).

Alternatively, the Cluster Analysis task can create non-hierarchical clusters of coordinate data by using the K-means method. You cannot use the K-means method with distance data. You can use this task to analyze population data.

For example, suppose that you want to determine whether national figures for birth rates, death rates, and infant death rates can be used to determine certain types or categories of countries. You can perform a cluster analysis to determine whether the observations can be formed into groups that are suggested by the data.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CLUSTER, FASTCLUS, TREE
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Cluster Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analyses variables	performs analyses using the variables that you assign to this role. You must assign at least one variable to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analysis by a variable that you have already assigned as **Set 1 variables** or **Set 2 variables**.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Identifying label	specifies the values to use to identify rows in the cluster history and in the output tree table. If no variable is assigned to this role, each row is denoted by OBS*n*, where *n* is the row number.
Copy variables	copies the values for these variable to the output tree data set. Observations in the output tree data set that represent clusters of more than one observation from the input data set have missing values for the copy variables.

Cluster Analysis: Setting Cluster Options

In the selection pane, click **Cluster** to access these options.

Option Name	Description
Cluster method	specifies the cluster methods to use in the analysis. ■ **Average linkage** calculates the group average by using the unweighted pair-group method which uses the arithmetic averages. ■ **Centroid method** uses the centroid method (unweighted pair-group method using centroids, UPGMC, centroid sorting, weighted-group method). ■ **K-means algorithm** uses Euclidean distances, so the cluster centers are based on least squares estimation. This method assumes that the cluster centers are the means of the observations that are assigned to each cluster when the algorithm is run to complete convergence. Each iteration reduces the least squares criterion until convergence is achieved. ■ **Ward's minimum variance method** uses Ward's minimum-variance method, which calculates the error sum of squares by using trace W.
K-means cluster options	specifies the additional options for the K-mean algorithm. ■ **Maximum number of clusters** specifies the maximum number of clusters that are allowed. ■ **Maximum number of iterations** specifies the maximum number of iterations for recomputing cluster seeds. ■ **Seed replacement** specifies how seed replacement is performed. □ **full** replaces the seed when the distance between the observation and the closest seed is greater than the minimum distance between seeds, or when the smallest distance from the observation to all seeds other than the nearest one is greater than the shortest distance from the nearest seed to all other seeds. □ **partial** replaces the seed only when the distance between the observation and the closest seed is greater than the minimum distance between seeds □ **none** suppresses seed replacement. □ **random** selects a simple pseudo-random sample of complete observations as initial cluster seeds. ■ **Specify random seed** specifies a positive integer as a starting value for the pseudo-random number generator. **Note:** This option is available only if you have selected random as the **Seed replacement** option.

Cluster Analysis: Generating Plots and Diagrams

In the selection pane, click **Plots** to access these options.

Note: Plots and diagrams are not available with the K-means algorithm cluster method.

A graphical view of the clustering process can often be helpful in interpreting the clusters. Plots are generated using a subset of the input data source. How the data is subset depends on whether the cubic clustering criterion (CCC) statistic can be computed. If the CCC statistic can be computed, then the data is subset based on the non-missing values of the CCC statistic. If the CCC statistic cannot be computed, then the data is subset based on the non-missing values of the pseudo *F*-statistic.

Option Name	Description
Pseudo F statistics	displays a scatter plot for the pseudo *F*-statistic against the number of clusters. **Note:** If you select the **Pseudo F and t-square statistics** option on the **Results** panel, then a graph of the pseudo *F* statistics is always generated.
Pseudo t-square statistics	displays a scatter plot for the pseudo t^2 statistic against the number of clusters. **Note:** If you select the **Pseudo F and t-square statistics** option on the **Results** panel, then a graph of the pseudo *t*-square statistics is always generated.
CCC statistics	displays a scatter plot for the cubic clustering criterion (CCC) statistic against the number of clusters. **Note:** This option is not available if the input data set is of type DISTANCE.
Tree diagram orientation	displays a tree diagram of the clusters. The objects that are clustered are leaves. The cluster that contains all objects is the root. A cluster that contains at least two objects but not all of them is a branch. The general term for leaves, branches, and roots is node. If a cluster A is the union of clusters B and C, then A is the parent of B and C, and B and C are children of A. A leaf is thus a node with no children, and a root is a node with no parent. If every cluster has at most two children, the tree diagram is a binary tree. Select Horizontal to orient the tree diagram with the height axis horizontal and the root at the left. Select Vertical to orient the tree diagram with the height axis vertical and the root at the top, which is the default.

Cluster Analysis: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Save output data	generates an output data set. Yyou can choose from these options: ■ Cluster tree data creates an output data set that can be used to draw a tree diagram of the cluster hierarchy. The data set contains one observation for each observation in the input data set, plus one observation for each cluster of two or more observations (that is, one observation for each node of the cluster tree). The total number of output observations is usually 2n-1, where n is the number of input observations. ■ **Tree output** creates an output data set for the average linkage method, centroid method, and Ward's minimum variance method. The data set contains one observation for each object in the tree and variables called CLUSTER and CLUSNAME that show cluster membership at any specified level in the tree. You can also specify the number of clusters or levels to include in the table. □ To specify the number of clusters to include in the output data set, select the **Cluster** check box. By default, the number of clusters is 10. However, the number of clusters might not equal the number that you specified if any of these conditions are met: ▪ The number of leaves in the tree is less than the number of desired clusters. ▪ The number of unconnected trees in the data set is larger than the number of desired clusters. ▪ A multiway tree does not contain a level with the specified number of clusters. □ To specify the number of levels to include in the output data set, select the **Prune** check box, and then specify the level of the tree that defines the disjoint clusters in the output data set. Only clusters between the root and the height of the specified level are displayed. ■ **K-means clusters** creates an output data set for the K-means algorithm method. The output data set contains all the original data, plus the CLUSTER and DISTANCE variables that are generated by the cluster analysis. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Option Name	Description
Display output	displays the output. You can specify these options: ■ **Display output** displays information about the history of the clustering process. This includes the number of clusters, the names of the clusters that are joined, and the number of observations in the new cluster. Use the Cluster generations box to specify the number of generations of the cluster history to display. ■ **Simple summary statistics** displays means, standard deviations, skewness, kurtosis, and a coefficient of bimodality. ■ **Cubic clustering criterion (CCC)** displays the cubic clustering criterion and approximate expected R-square under the uniform null hypothesis. This option applies only to coordinate data. These statistics are useful in determining the number of clusters in the data. Values of CCC greater than 2 or 3 indicate good clusters; values between 0 and 2 indicate potential clusters, but they should be considered with caution; large negative values can indicate outliers. Note: This option is not available if the input data set is of type DISTANCE. ■ **Pseudo F and t-square statistics** displays pseudo F and t-square statistics. These statistics can be useful indicators of the number of clusters. Relatively large values of the pseudo F statistic indicate a stopping point. A general rule for interpreting the values of the pseudo t-square statistic is to move down the column until you find the first value that is markedly larger than the previous value and then to move back up the column by one cluster. Note: When you select this option, plots of the pseudo *F* and *t*-square statistics are generated whether you select the Pseudo **F** statistics and Pseudo *t*-square statistics options on the **Plots** panel. Note: **Display output** is the only display option available with the K-means algorithm cluster method.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

18

Compare Data

About the Compare Data Wizard

The Compare Data wizard helps you create a report that compares two data sets or compares two variables within or across data sets.

You can use the Compare Data wizard to compare the changes that have been made to a SAS data set. For example, suppose you have copied a data set, and you want to see if there are any differences between your copy and the original data set. The Compare Data wizard enables you to view a summary of the changes that have been made to the original data set.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	COMPARE
Required SAS products	Base SAS
Recommended additional SAS products	None

Compare Data: Select the Data

1. Specify whether you want to compare variables between two data sets or compare variables within the same data set.
2. Select the base and comparison data sets that you want to use.
3. Click **Next** to select the variables to compare.

Compare Data: Select the Variables to Compare

By default, all variables that have the same name and type are matched in the table. You can include additional variables, edit the variables to compare, and delete variables.

To specify the variables to compare:

1. Before you can select the variables to compare, you must select the data.
2. From the **Match data by** drop-down list, select the criteria for comparing the variables. You can choose to match the data either by observations or by ID variables. If you choose to match the data by ID variables, then the variable must exist in both the base and comparison data sets. If you are comparing variables within the same data set, then you can match the data only by observations.
3. To add a new variable pair to the table, click **New**. A new row is added to the table. Use the drop-down lists in each cell to select the variables to compare.

   If one of the variable pairs is incorrect, then use the drop-down list in the cell to select the correct variable name. You can also click **Edit**.

   To delete a variable pair, select the row from the table and click **Delete**.
4. Click **Next** to specify the comparison criteria.

Compare Data: Specify the Comparison Criteria

1. Before you can specify the comparison criteria, you must select the variables to compare.
2. Specify the value of the equality criterion in the **Equality Criterion** box.
3. Select the method of determining equality from the **Method for Judging Equality** drop-down list. Numeric values are determined to be unequal if the magnitude of their difference, as determined by the method that you select, is greater than the value of the equality criterion. The following methods are available:
   - **Absolute** - compares the absolute difference of the values to the value of the equality criterion. If the absolute value of y minus x is greater than the equality criterion, then the values are determined to be unequal.
   - **Exact** - tests for exact equality. If the value of y does not equal the value of x, then the values are determined to be unequal.
   - **Percent** - compares the absolute percent difference to the value of the equality criterion.
   - **Relative** - compares the absolute relative difference to the value of the equality criterion.
4. Specify how to treat missing values. You can choose from these options: **Treat a missing value in the BASE data set as equal to any value** or **Treat a missing value in the COMPARISON data set as equal to any value**.
5. Click **Next** to specify the output options.

Compare Data: Specify the Output Options

By default, the Compare Data task creates an HTML report with a data set summary, a variables summary, an observation summary, a values comparison summary, and the value comparison results. You can also choose to create an output data set that contains observational output and an output data set that contains the summary statistics.

1 Before you can specify the output options, you must specify the comparison criteria.

2 Select **Include output data set** to create an output data set that contains a row for each matching observation. This data set contains a column for each variable in the observation, a column for the type of observation (_TYPE_), and a column for observation number (_OBS_). The values in the _TYPE_ column can be one of these types:

- BASE - The values in this observation are from an observation in the base data set.
- COMPARE - The values in this observation are from an observation in the comparison data set.
- DIF - The values in this observation are the differences between the values in the base and comparison data sets.
- PERCENT - The values in this observation are the percent differences between the values in the base and comparison data sets.

You can choose to include these observations in the output data set:

- **Write an observation for each observation in the base data** writes in the output data set the observations in the base data set. The value in the _TYPE_ column of the output data set is set to BASE.
- **Write an observation for each observation in the comparison data** writes in the output data set the observations in the comparison data set. The value in the _TYPE_ column of the output data set is set to COMP.
- **Include difference value** writes in the output data set the differences between the values in the base and comparison data sets. The value in the _TYPE_ column of the output data set is set to DIF.
- **Include values for the percent differences** writes in the output data set the percent differences between the values in the base and comparison data sets. The value in the _TYPE_ column of the output data set is set to PERCENT.
- **Suppress observations when all values are equal** does not include in the output data set any observations where all of the values of the variables are equal.

3 Select **Include output data set for summary statistics** to create an output data set that contains a row for each summary statistic for each pair of variables. The output data set contains these columns:

- _VAR_ - contains the name of the variable from the base data set.
- _WITH_ - contains the name of the variable from the comparison data set.
- _TYPE_ - contains the name of the statistic in the observation.
- _BASE_ - contains the value of the statistic that is calculated from the values of the variable in the base data set with matching observations in the comparison data set.
- _COMP_ - contains the value of the statistic that is calculated from the values of the variable in the comparison data set with matching observations in the base data set.
- _DIF_ - contains the value of the statistic that is calculated from the differences of the values of the variable in the base data set and the matching variable in the comparison data set.

- _PCTDIF_ - contains the value of the statistic that is calculated from the percent differences of the values of the variable in the base data set and the matching variable in the comparison data set.

4 Specify the maximum number of differences to print.

5 Click **Finish** to generate the output.

19

Contour Plot

About the Contour Plot Task

The Contour Plot task creates line, filled, pattern, or smooth plots that show the mathematical relationships between three numeric variables. The plots represent three-dimensional relationships in two dimensions. Lines or areas in a contour plot represent levels of magnitude (z) that correspond to a position (x, y) on a plane.

You might want to use this task to illustrate an occurrence that is found in nature. For example, you can create a contour plot that illustrates the percentage of clay that is found in soil samples at various locations of a testing site. The X and Y axes on the plot represent a graph of surface height at various x-y locations. The contour lines within the plot represent the locations on the plane that have the clay percentages that are specified in the legend.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GCONTOUR, G3GRID
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Contour Plot: Selecting a Plot Type

In the selection pane, click **Contour Plot** to access these options.

Option Name	Description
Contour Line Plot	uses lines to represent the Z (or contour) variable.
Filled Contour Plot	fills each of the contour lines with a different color. One color specifies each level of contour.
Filled Outline Contour Plot	fills each of the contour lines with a different color and then outlines each area to clearly indicate where the contour lines are.
Pattern Contour Plot	represents the contour levels with rectangles filled with patterns. The pattern for each rectangle is determined by calculating the mean of the values of the Z variable for the four corners of the rectangle and assigning the pattern for the level closest to the mean.
Smooth Contour Plot	combines adjacent grid cells with the same pattern to form a single pattern area.

Contour Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart.
Depth	The column that you assign to this role is the third-dimension or Z-axis variable for the plot.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value.

Contour Plot: Setting Appearance Options

Setting Contour Plot Options

In the selection pane under the Appearance heading, click **Contour** to access these options.

Role Name	Description
Specify the number of contour levels	Specify the number of contour levels—that is, the number of regions in which to divide the values of the variable that is assigned to the **Depth** role (or Z-axis variable).
Automatically generate labels for contour levels	Select the check box to automatically label the contour lines. The label for each contour line is the **Depth** value for that contour level. The legend will still appear. **Note:** This option is not supported by the ActiveX, ActiveX image, Java, and Java image graph formats.
Preprocess data with interpolation options	If you run the Contour Plot task and receive an error in the log that "Less than half of the grid cells have data values," select this check box before you rerun the task. The software creates a data set whose horizontal (X and Y) variable values form a complete grid, and it interpolates the value of the vertical (Z) variables for each point on the X-Y plane. You can control the interpolation method by selecting one of these interpolation options: ■ **Join** to use a linear interpolation within a set of triangular regions. This interpolation method creates values in the range of the initial values of the vertical variable, but the resulting interpolated surface might not be smooth. ■ **No scaling** to prevent scaling of the X and Y variables to the same range before interpolation. ■ **Partial** to use a spline to estimate the derivatives for the interpolation. A bivariate spline is fit to the nearest neighbors and used to estimate the needed derivatives. The contours are not as smooth as the Spline option, but fewer computer resources are used. ■ **Spline** to use a bivariate spline for the interpolation. This method can be time-consuming, especially if there are more than 100 data points.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.

Role Name	Description
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. Note: This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Note: If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.

Role Name	Description
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.

Role Name	Description
Chart background color	specifies the background color. Note: If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. Note: This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. Note: The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Setting Advanced Options

In the selection pane under the Appearance heading, click **Advanced** to access these options.

Role Name	Description
Statistic used to calculate bar	specifies how to calculate the length of the bar. The available statistics depend on the graph that you are creating. ■ If you did not assign a variable to the **Sum of** role in the Bar Chart task or **Bar sum of** role in the Bar-Line Chart task, then the frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. ■ If you assigned a variable to the **Sum of** or **Bar sum of** role, or if you selected multiple measures group as the type of bar chart, then the sum and average statistics are available.
Statistic used when plotting line	specifies the statistic to use when plotting the line on a bar-line chart. The frequency, cumulative frequency, percentage, and cumulative percentage statistics are available. If you assigned a variable to the **Line sum of** role in the Bar-Line Chart task, then this option is not available from the **Advanced** panel. Instead, this option is available from the **Plots** panel, so you can control the statistic for each line when more than one variable is assigned to the **Line sum of** role.

Role Name	Description
Display error bars	displays confidence intervals on the bars. From the **Type** drop-down list, select from these options: ■ **Bars** draws error bars as bars half the width of the main bars. ■ **Both** draws error bars as two tick marks that are joined by a line. ■ **Top** draws the error bar as a tick mark for the upper confidence limit that is joined to the top of the bar by a line. Select a color for the error bars, and use the **CLM** box to specify a confidence level. **Note:** You can display error bars only when these are met: ■ you select **Average** or **Percentage** as the statistic that is used to calculate the length of the bars ■ you are not creating a three-dimensional grouped bar chart ■ you have not assigned a variable to the **Stack** role
Accept missing values	specifies to accept missing values as a valid value for the chart column.
Skip zero values	suppresses any values of the chart column for which the bar length is zero.
Calculate percentages and cumulative percentages for each group (G100)	calculates percentages and cumulative percentages separately for each group in grouped bar charts. When you select this option, the individual percentages reflect the contribution of the midpoint to the group; they total 100 percent for each group. **Note:** This option is available only when you assign a variable to the **Group bars by** role for grouped bar charts.
Show statistics next to bar	shows the value of the statistic that is used to calculate the bar. **Note:** This option is available only for horizontal bar charts.
Specify one statistical value to show for bars	shows the value of a statistic that you select from the drop-down list. The statistic can be the same as or different from the statistic that is used to calculate the bar length. If you are creating a two-dimensional bar chart, you can choose to display the value of the statistic inside the bar.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

20

Copy Files

About the Copy Files Task

The Copy Files task enables you to transfer files from your local computer to a SAS server. You can also use this task to copy files from the SAS server to your local computer. The Copy Files task works in a similar way to an FTP application. However, this task relies on the SAS protocols to complete the file transfers and does not require an FTP server.

This task uses SAS Integration Technologies to copy the file data across the network. The task does not rely on PROC DOWNLOAD or PROC UPLOAD from SAS/CONNECT.

Note: This task is available only in SAS Enterprise Guide. You cannot run the code generated by this task outside of SAS Enterprise Guide. If you export a SAS Enterprise Guide project or a process flow as a SAS program, any steps for the Copy Files tasks are not included.

Copy Files from Your Local Computer to a SAS Server

1. Select the server to use as the source or destination of the files that you want to copy.
2. Select the type of transfer. You can upload files from your local computer to the SAS server or download files from the SAS server to your local computer.
3. Specify the full path of the files that you want to copy. In the **Source files to copy** box, you can use standard wildcard characters to match multiple files. An asterisk (*) matches all characters, in any number, before the next non-wildcard character. A question mark (?) matches any single character that occupies that position in the filename.

   Here are some examples of wildcards and the patterns that they match: Note

   `/u/dept/finance/*.xls`
   Matches all of the XLS files within the /u/dept/finance folder.

   `C:\Data\??Sep2012.csv`
   Matches all files whose names begin with two characters and end with "Sep2012.csv". Example matches include "01Sep2012.csv" and "15Sep2012.csv". The filename "199Sep2011.csv" would not match this pattern, because the prefix is three characters, not two.

`C:\Data\15???2012.*`
Matches all files whose names begin with "15" followed by any three characters, followed by "2012", followed by any file extension. Example matches include "15Sep2012.csv" and "15Oct2012.xls".

Note: Wildcard characters are permitted only in the filename and not within the folder path. For example, `c:\data\Sept*\report.xls` is not a valid source file specification for this task.

4 Specify where you want the files to be copied on your local computer or SAS server. When specifying a destination folder, remember these rules:

- The folder must already exist on the target system. The Copy Files task does not create new folders.
- You must have appropriate file permissions to write content to the folder.
- You must specify an absolute path for the target folder (even when uploading files and the SAS workspace is configured with a starting root path of SASUSER or another non-system-root path).

5 Specify whether you want to use macro variables in the file or folder specifications. Select the **Resolve SAS macro variables in source and destination paths** check box to make the Copy Files task behave more dynamically by using SAS macro variables and expressions in the filename or the destination folder path. When this option is selected, the Copy Files task attempts to evaluate the macro variables and macro functions within these fields before beginning the copy operation.

Here are some examples of values that contain macro references and how they might resolve.

`/u/&SYSUSERID/Data`
Resolves to a folder that contains the user ID of the SAS user within the SAS Workspace Server.

`C:\Data\%sysfunc(substr(&SYSUSERID,,3))`
Resolves to a folder that contains only a portion of the SAS user ID within the SAS Workspace Server.

`/u/Data/&outputFile..xls`
Resolves to an XLS filename where the root of the filename is stored in the &OUTPUTFILE macro variable. Using this macro variable assumes that a previous process (perhaps a SAS program) has set the macro variable to match an existing filename.

CAUTION! The macro variables must exist within your SAS session (either as SAS built-in macro variables or defined in an earlier process). If the macro variables are not specified correctly or if you use incorrect syntax for a macro function, the Copy Files task can generate errors or unexpected results.

To verify that the macro expressions resolve correctly, click **Test macro values**. A short SAS program runs to resolve the macro expressions, and the results are displayed in a new window. Remember that when you test these values, the macro expressions resolve to values that reflect the current state of the macro variables or the results of any macro functions that you use. These values can be different when you actually run the task later.

6 Specify whether to overwrite any existing files in the destination folder. By default, the Copy Files task does not overwrite any existing files. If a source file matches the name of an existing file, the task does not copy that file and adds a note to the log.

7 Specify whether the task automatically detects and rewrites (using the appropriate line-ending style) any text files that are transferred. UNIX and Windows environments use different conventions for the line-ending characters in text files. UNIX environments use a line feed (LF) character, and Windows environments use a line feed (LF) followed by a carriage return (CR). The absence of the CR character can make UNIX text files difficult to read in standard text editing applications such as Notepad.

21

Correlations

About the Correlations Task

Data correlation is a statistical procedure for describing the relationship between numeric variables. The relationship is described by calculating correlation coefficients for the variables.

You might want to use the Correlations task to learn more about product sales. For example, you can use this task to determine whether there is a correlation between a customer's size (as measured by sales) and the amount of product that the customer purchases from your company.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CORR
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Correlation: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	performs analyses using the variables that you assign to this role. You must assign at least one variable to this role.
Correlate with	correlates each variable assigned to this role with the analysis variables.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. Note: To prevent the input data set from being sorted, clear the **Sort by variables** check box. Note: You cannot group analysis by a variable that you have already assigned as **Analysis variables** or **Correlate with**.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Partial variables	specifies the variables to use in the Pearson's partial correlation, Spearman's partial rank-order correlation, or Kendall's partial tau-b, depending on the type of correlation that you choose.
Relative weight	specifies the value of the variable in each row to calculate product-moment correlation coefficients. You can assign a maximum of one column to this role.

Correlation: Setting Correlation Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Correlation types	specifies the correlation analysis. ■ **Pearson** calculates Pearson product-moment correlation. This is a parametric measure of association for two continuous random variables. The correlations range from -1 to 1. ■ **Hoeffding** calculates Hoeffding's measure of dependence, D. This is a nonparametric measure of association that detects more general departures from independence. This D statistic is 30 times larger than the usual definition and scales the range between -0.5 and 1 so that only large positive values indicate dependence. ■ **Kendall** calculates Kendall tau-b. This is a nonparametric measure of association that is 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. Kendall's tau-b ranges from -1 to 1. ■ Spearman calculates Spearman rank-order correlation. This is a nonparametric measure of association that is based on the rank of the data values. The correlations range from -1 to 1. Note: The Hoeffding, Kendall, and Spearman correlation types are unavailable if you assign a variable to the **Relative weight** role. The Hoeffding correlation type is also unavailable if you assign a variable to the **Partial variables** role.
Pearson correlations options	specifies the options for the Pearson correlation: ■ **Cronbach's coefficient alpha** calculates Cronbach's coefficient alpha. Separate coefficients that use raw and standardized values (by scaling the variables to a unit variance of 1) are computed. For each correlation variable, the correlation between the variable and the total of the remaining variables is computed. Cronbach's coefficient alpha is computed by using only the remaining variables. Note: This option is unavailable if you assign a variable to the **Correlate with** role. ■ **Covariances** calculates a covariance matrix. If you assign a partial variable, then a partial covariance matrix is computed. •Sums of squares and cross products calculates the sum of squares and crossproducts. If you assign a partial variable, then the unpartial SSCP matrix is computed. ■ **Corrected sums of squares and cross products** calculates the corrected sums of squares and crossproducts. If you assign a partial variable, then both an unpartial and a partial CSSCP matrix are computed. ■ **Suppress Pearson correlations from results** excludes Pearson correlations in the results.

Option Name	Description
Fisher correlation options	uses the Fisher options to request confidence limits and *p*-values under a specified null hypothesis for correlation coefficients using Fisher's *z* transformation. Note: These options are available if you select **Pearson** or **Spearman** as the correlation type. ■ **Specify the value of alpha** specifies the level of the confidence limits for the correlation. The value of this option must be between `0` and `1`. The default value is `0.05`. ■ **Type of confidence limits** specifies the type of confidence limits. The default is a two-sided confidence limit from the two-sided alternative. ■ **Specify the value rho0 in the null hypothesis** specifies the value of rho in the null hypothesis. The default value is `0`. ■ **Use bias adjustment for constructing confidence levels** specifies whether the bias adjustment is used in constructing confidence limits. By default, this check box is selected, and a new correlation estimate is produced by using the bias adjustment.
Divisor for variance	specifies the divisor that is used in the calculation of variances, standard deviations, and covariances. By default, the divisor for the variance is the degrees of freedom.
Omit rows with missing values being correlated	uses only those rows in which none of the analysis variables contain missing values. By default, the Correlations task uses all nonmissing pairs of values for each pair of analysis variables. This means that some correlations are computed by using more observations than others.

Correlation: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Plots	generates a scatter plot for each correlation pair.

Option Name	Description
Results to display	specifies the options for the correlation analysis. In the results, you can include these values: ■ show summary statistics for each analysis variable, including mean, standard deviation, sum, minimum, and maximum. ■ show the significance probabilities that are associated with each correlation. ■ rank the correlation coefficients from highest to lowest. For each analysis variable, the **Correlate with** variables are listed in order of descending absolute value of correlation coefficient. By default, the correlation coefficients are not ranked. If you choose to rank the correlation coefficients, you can also specify how many correlations to show for each variable. If you specify fewer correlations than the maximum number, the Correlations task ranks the correlations from highest to lowest in absolute value and reports the best n correlations, where n is the number of correlations that you specified.

Correlations: Create the Output Data

In the selection pane, click **Output Data** to access these options.

Select the **Save output data** check box. The correlation statistics that are included in the final output depend on the types of correlations that are selected in the **Correlation types** area.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

22

Create Format

About the Create Format Task

The Create Format task creates a numeric or character format. A numeric format specifies the text that will appear when the stored values of a numeric variable are displayed or printed. A character format does the same for a character variable. You can use this task to format your output.

You can use this task to these tasks:

- print numeric values as character values (for example, print 1 as MALE and 2 as FEMALE)
- print one character string as a different character string (for example, print YES as OUI)
- print numeric values by using a template (for example, print 9458763450 as 945-876-3450).

Note: This task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	FORMAT
Required SAS products	Base SAS
Recommended additional SAS products	None

Create Format: Setting Format Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Format name	specifies the name for the format that you are creating. This name will be used to refer to the format each time you apply it. Format names must comply with these rules: ■ They can contain only letters, digits, and underscores; all other characters are prohibited. GENDER and M_OR_F are valid format names, but M-OR-F and M#F are not valid. ■ They cannot begin or end with a digit. HEX2DEC is a valid format name, but 4MAT is not valid. ■ They have a maximum length that depends on the format type. The name of a character format cannot be longer than 31 characters. The name of a numeric format cannot be longer than 32 characters. **Note:** If you running a version of SAS prior to SAS 9, then the name of the character format cannot be longer than 7 characters, and the name of the numeric format cannot be longer than 8 characters. Using a longer format name can cause unexpected results. MALEORFEMALE is too long to be a valid format name of either type. STATCODE is a valid name for a numeric format, but it is too long to be the name of a character format.
Specify format width	specifies the width of the format. By default, the format width is the length of the longest formatted value.
Location to store format	specifies the server and library where you want to store the new format. By default, the format will be stored in the Egtask library on the selected server, if there is one. If there is no Egtask library, then the format will be stored in the Sasuser library. In either case, the format will remain in existence when you exit and will be available for future use. **Note:** If you choose to store the format in the Work library, then the format will be deleted when you exit. A format can exist and yet be unavailable for use. A format is available for use only if the library in which it is stored is included in the format search path.
Format type	specify whether you are creating a character format or a numeric format. To protect against the inherent imprecision of floating-point storage, a numeric format is applied with a fuzz factor. By default, this fuzz factor is `1E-12`. To change the fuzz factor, select the Specify fuzz factor check box and specify a new value in the box.

Create Format: Defining the New Format

How to Define a New Format

In the selection pane, click **Define Formats** to access this page.

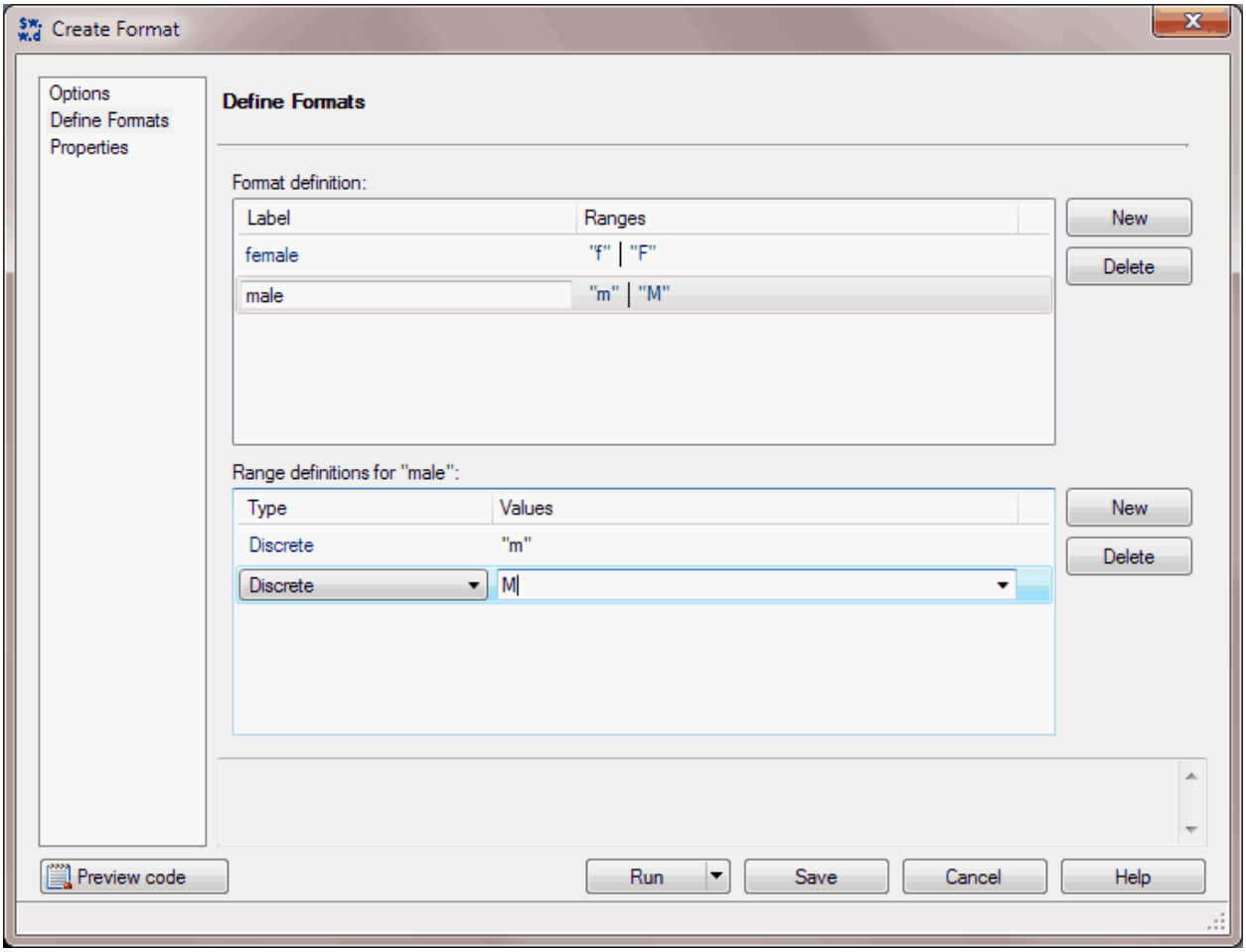
Use this page to map the stored values of a variable to the text that will appear when you display or print those values. There are two parts to the format definition. The **Label** column specifies the text that will be displayed or printed, and the **Ranges** column specifies one or more ranges of stored values that will be translated into that text. A red box around a range indicates a conflict between that range and another range. For example, a conflict occurs if you map all numbers between 0 and 5 to the label LOW and all numbers between 3 and 7 to the label MEDIUM.

To define a new format

1. Next to the **Format definition** box, click **New**. A new row is added to the **Format definition** and **Range definitions** boxes.
2. In the **Label** column, enter the text to be displayed or printed
3. In the **Range definitions** box, specify the stored values that will be mapped to that text. To specify a range definition:
   a. Select a type for the range definition (either **Discrete** or **Range**).
   b. Enter the values to map to the label in the format definition. Each discrete value or range must be entered as a separate item.

Example 1: Character Format

You are defining a character format for a variable named GENDER that has the following four stored values: f, F, m, M. You want the software to display "female" for the stored values f and F, and "male" for the stored values m and M. The **Define Formats** page should look similar to this display:

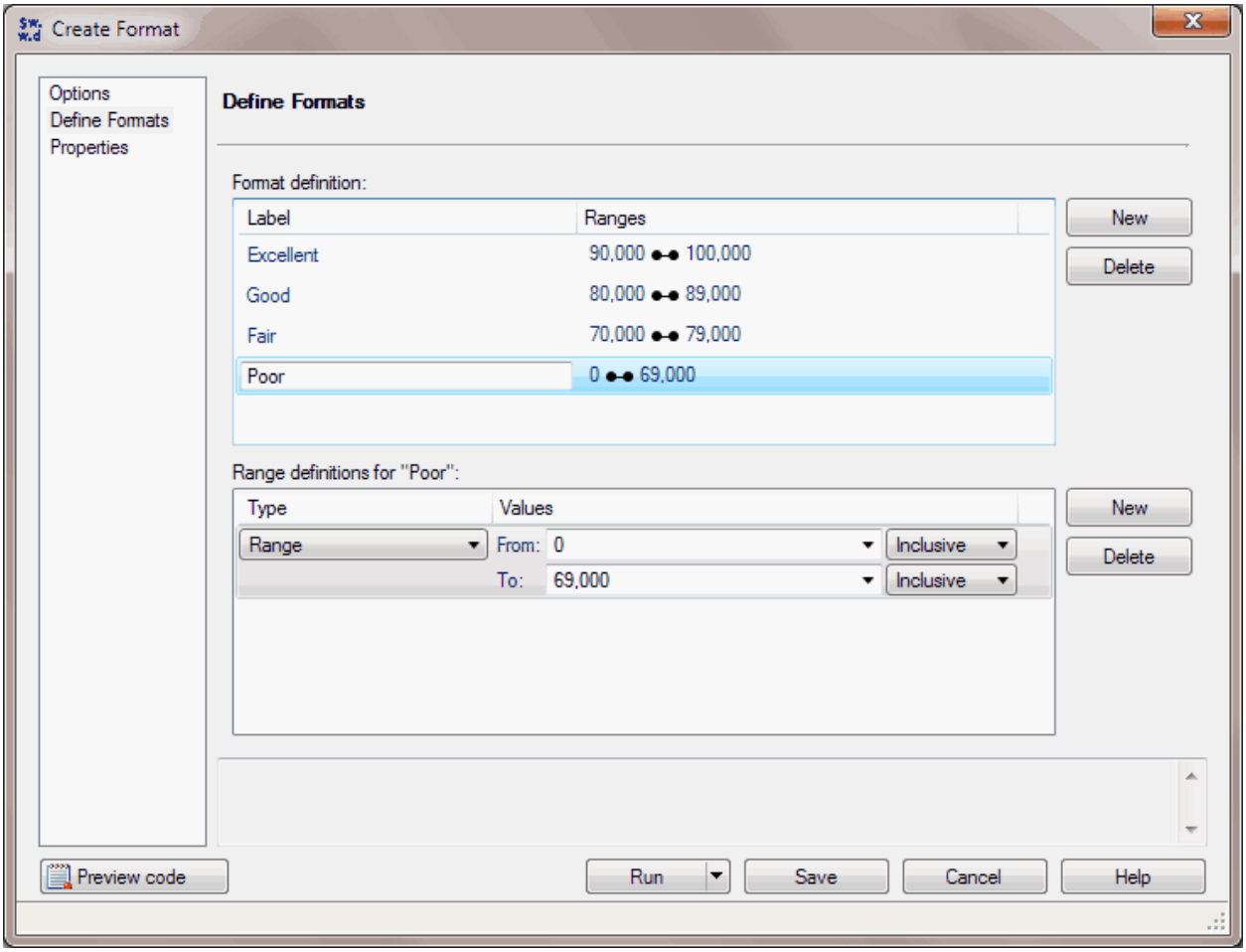


Example 2: Numeric Format

You are defining a numeric format for a variable named SALES. You want the software to display a rating based on the range of sales values.

- Excellent = 90,000–100,000 (inclusive)
- Good = 80,000–89,000 (inclusive)
- Fair = 70,000–79,000 (inclusive)
- Poor = 0–69,000 (inclusive)

The **Define Formats** page should look similar to this display:



Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

23

Create Format from Data Set

About the Create Format from Data Set Task

The Create Format from Data Set task enables you to create a SAS format by using data that is saved in a SAS data set. A SAS format enables you to map raw data values to a formatted value. You can create SAS formats for character or numeric data.

Here are examples of when you might want to use the Create Format from Data Set task:

- Next year, your organization plans to conduct customer focus groups in each region (North, South, East, West) of the United States. Your 2009 sales data lists the address and state for each customer. However, you want this sales data categorized by region. In the REGIONS data set, you have already mapped the states to their corresponding regions:

State	Region
Arizona	West
California	West
Florida	South
Vermont	North
Virginia	East

  Using the REGIONS data set as input in the Create Format from Data Set task, you can create a SAS format that you can later apply to the state values in your sales data.

- Recently, your company conducted several clinical trials. The patient data includes the age of each trial participant. In the AGES data set, you have already divided the age ranges into discrete categories:

Age	Low	High
Infant	0	2
Toddler	3	4

Age	Low	High
Child	5	9
Preteen	10	12
Teenage	13	17
Adult	18	100

Note: This task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	FORMAT
Required SAS products	Base SAS
Recommended additional SAS products	None

Create a Format from a Data Set

1. Select the input data set. By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Source data set** field.
2. Specify a name for the format.

   Format names must comply with the following rules:

   - They can contain only letters, digits, and underscores; all other characters are prohibited. GENDER and M_OR_F are valid format names, but M-OR-F and M#F are not valid.
   - They cannot begin or end with a digit. HEX2DEC is a valid format name, but 4MAT is not valid.
   - They have a maximum length that depends on the format type. The name of a character format cannot be longer than 31 characters. The name of a numeric format cannot be longer than 32 characters.

   Note: If you are running a version of SAS prior to SAS 9, then the name of the character format cannot be longer than 7 characters, and the name of the numeric format cannot be longer than 8 characters. Using a longer format name can cause unexpected results. MALEORFEMALE is too long to be a valid format name of either type. STATCODE is a valid name for a numeric format, but it is too long to be the name of a character format.
3. Specify whether you are creating a character format or a numeric format. To protect against the inherent imprecision of floating-point storage, a numeric format is applied with a fuzz factor. By default, this fuzz factor is `1E-12`.
4. Select the library where you want to store the new format. By default, the format is stored in the Work library on the selected server. However, if you store the format in the Work library, then the format is deleted when you exit SAS Enterprise Guide. To make the format available for future use, select a different output library.

Note: A format can exist and yet be unavailable for use. A format is available for use only if the library in which it is stored is included in the format search path.

5 Specify whether the values in the input data set are discrete values or a range of values.

To create a format from discrete values:

a In the **Value types** box, select **Discrete/Lookup**.

b In the **Variables** area from the **Discrete values** drop-down list, select the variable in the data set that contains the values that you want to format. The values for the selected variable appear in the **Look Up Value** column in the preview table.

c From the **Labels** drop-down list, select the variable in the data set that contains the labels for the values. The values for the selected variable appear in the **Label** column in the preview table.

To create a format from a range of values:

a In the **Value types** box, select **Ranges**.

b In the **Variables** area from the **Low values** drop-down list, select the variable in the data set that contains the starting values for the range. The values for the selected variable appear in the **Start** column in the preview table.

c From the **High values** drop-down list, select the variable in the data set that contains the ending values for the range. The values for the selected variable appear in the **End** column in the preview table.

d From the **Labels** drop-down list, select the variable in the data set that contains the labels for the values. The values for the selected variable appear in the **Label** column in the preview table.

6 Specify the maximum length for the label. By default, the label can be eight characters in length.

7 Specify a label for other values. When applying this format to a data source, SAS Enterprise Guide might find that some values in the data are not accounted for in the format. You can create a unique label for any values in the data source that do not have a corresponding formatted value. By default, the label for these values is **Other**.

8 Specify whether to include a summary report in your results. By default, a summary report is not created when you run the task.

24

Create Map Feature Table

About the Create Map Feature Table Task

The Create Map Feature Table task creates a feature table for user-defined or other map data sets that you use with the Map Graph task. It also enables you to project a map data set—that is, it converts the spherical coordinates in a map data set to Cartesian coordinates. This task enables you to use one of several map projection techniques to project the coordinates in a traditional map data set onto a two-dimensional plane while attempting to minimize the distortion area, distance, direction, and shape properties of the original sphere.

Note: This task requires the %ADDFEAT macro, which is shipped with SAS Release 8.2 and later releases. Otherwise, contact SAS Technical Support for instructions on obtaining this macro.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GPROJECT
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	MAPS

Create Map Feature Table: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

After you have selected a data source, you can assign variables to roles.

Note: The input data source for this task is required to have an X column and a Y column, which contain the spatial data coordinates for the map. The Maps library contains map data sets that already have an associated feature table. For example, the US2 data set is the feature table for the U.S. (United States) map.

The Maps library is protected, so you cannot use the data sets in this library as input data for the Create Map Feature Table task. However, you can copy data sets from the Maps library to an unprotected location, and then use these data sets in this task.

Role Name	Description
Identification variables	The variables that you assign to this role are used to identify sections within the map. At least one identification variable is required to both project the data and/or create a feature table.
Attribute variables to keep	These are variables in the input map (geometry) data set that you also want to include in the feature table that you will create. **Note:** This data role appears only if the **Create a map feature table** option is selected.
Longitude variable to project	This is the variable in the input data set that contains the unprojected (spherical) longitude data point. **Note:** This data role appears only if the **Project the map data (from spherical to Cartesian coordinates)** option is selected.
Latitude variable to project	This is the variable in the input data set that contains the unprojected (spherical) latitude data point. **Note:** This data role appears only if the **Project the map data (from spherical to Cartesian coordinates)** option is selected.

Create Map Feature Table: Setting the Options

Setting Map Data Projection Options

In the selection pane, click **Options** to access these options.

1. Select the **Project the map data (from spherical to Cartesian coordinates)** check box.
2. Select the appropriate projection method from the **Projection method** drop-down list.
   - **Albers' equal-area projection** is a conic projection from the surface of the sphere to a cone secant to the sphere, cutting it at two standard parallels of latitude. The axis of the cone coincides with an extension of the polar axis of the sphere. Each section of the resulting map bears a constant ratio to the area of the sphere. In general, distortion in shape tends to increase toward the poles in latitudes outside of the two standard parallels. Albers' projection is well suited to portray areas of large and small east-to-west extent. However, both standard parallels must lie on the same side of the equator.
   - **Lambert's conformal projection** is a projection in which the meridians of longitude are straight lines that radiate from the apex of a secant cone and the parallels of latitude are concentric circles. It is ideal for navigational charts and maps of relatively small east-to-west extent. However, both standard parallels must lie on the same side of the equator.
   - **Gnomonic projection (azimuthal)** is a planar projection from the surface of the sphere directly onto an imaginary plane tangent to the sphere at the map projection pole. By default, the projection pole is placed

at the center of the map data set that is to be projected. Distortion increases as the distance from the map pole increases. This projection is best suited for mapping areas of small east-to-west extent.

3 Select the **Coordinate units** for the unprojected longitude and latitude values in the input map data set. The default is radians.

4 Select the **Direction of increasing longitude**. The default is West.

5 If you want to retain observations for which the projected values of the X and Y variables are identical to those in the previous observation, select the **Retain successive identical observations** check box. By default, successive identical observations are deleted.

Setting the Feature Table Options

In the selection pane, click **Options** to access these options.

To create a new feature table associated with the input map data set:

1 Select the **Create a map feature table** check box.

2 If you want to change the name of the map geometry variable, then specify a new name in the **Name of new map geometry variable** box. The name cannot start with a number or contain any spaces, punctuation marks, or special characters.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created

- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

25

Create Time Series Data Wizard

About the Summary Tables Wizard

The Create Time Series Data wizard helps you convert transactional data into fixed-interval time series. Transactional data is time-stamped data that is collected over time with irregular or varied frequency.

For example, you have data for sales from your company's website. These sales are time-stamped but not collected at a particular interval. You can use this task to convert the data into a time series data set.

Create Time Series Data Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to assign variables to roles.

Create Time Series Data Wizard: Assign Variables to Roles

1 Before you can assign variables to roles, you must select the data.

2 Assign variables to these required roles:

- **Time ID** - The column that you assign to this role is the time ID for the data. The time ID variable is a variable in the input data set that contains the SAS date, time, or datetime value of each observation. This variable is used to determine the frequency and ordering of the data and to extrapolate the time ID values

for the forecasts. You can assign only one variable to this role, and it must be either a date variable, a datetime variable, or a numeric variable that contains date or datetime values.

- **Time series for** - The column that you assign to this role is the variable to model and forecast. You must assign at least one numeric variable to this role. Use the up and down arrows to specify the order of the variables.

3 (Optional) Assign a variable to the **Separate time series for values of (Group variable)** role. This variable groups together observations that have the same value for the group-by variable. Assigning a group-by variable enables you to obtain separate analyses for groups of observations. You can assign character and numeric variables to this role.

The order of the BY variables describes the structure of the hierarchy. Use the up and down arrows to specify the order of the variables.

4 Click **Next** to specify the options for accumulation, missing values, and zero values.

Create Time Series Data Wizard: Specify the Options for Accumulation, Missing Values, and Zero Values

1 Before you can specify these options, you must assign variables to roles.

2 If the task has detected the interval from the input data set, then the default interval appears in the **Interval** drop-down list. You can change the interval, but the drop-down list contains only the intervals that are appropriate for the variable that you assigned to the **Time ID** role.

For large data sources, the task might take longer to detect an interval. If the interval has not been detected by the time at which you get to this step, you can choose an interval from the **Interval** drop-down list. Until an interval is detected, all possible values for the interval options are listed. After the interval is determined, if the type of variable that you assigned to the Time ID role and the interval value are in conflict, a warning message appears. The list of interval values is changed to only the values that are appropriate for the variable type. For more information about each statistic, see the Glossary of Statistical Terms.

3 Select the time series variable from the **Time Series** pane. These are the variables that you assigned to the **Time series for** role.

4 Specify the statistic for aggregating data across the levels of the hierarchy from the **Accumulation Statistics** drop-down list.

5 Specify how to treat missing values from the **Treat missing values as** drop-down list.

6 Repeat steps 3–5 for each time series variable.

7 Click **Next** to specify the output options.

Create Time Series Data Wizard: Select Additional Output

1 Before you can specify the output options, you must specify the options for accumulation, missing values, and zero values.

2 Select the **Time series graph** check box if you want to include a graph of the time series data in the output.

3 Select the **Save time series to a data set** check box.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

4 Click **Next** to specify the titles and footnotes.

Create Time Series Data Wizard: Specify the Title and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1 Before you can specify the title and footnotes, you must specify the output options.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

3 Click **Finish** to run the task.

26

Data Set Attributes

About the Data Set Attributes Wizard

The Data Set Attributes wizard helps you create a report with the data set's creation date, location, and number of observations as well as the variable names, labels, types, and formats.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	DATASETS
Required SAS products	Base SAS
Recommended additional SAS products	None

Data Set Attributes: Select the Data

1 Click **Add**. The Open dialog box appears.

2 Select the data set(s) that you want to use. To select multiple data sets, press the Ctrl key. Click **Open**. The selected data sets appear in the **Data** box.

3 Click **Next** to specify the output.

Data Set Attributes: Specify the Output

1 Before you can specify the output options, you must select the data.

2 Specify the location of the output data set. By default, the output data set is saved in the Sasuser library. To change the location or name of the output data set, click **Browse**. The resulting data set is also added to the project.

3 Select the output type. You can choose from these options:

- **Default Report** contains the output from the DATASETS procedure. This report includes the following data attributes: the date on which the data set was created and last modified, the number of observations, the encoding, any engine/host-dependent information, and an alphabetic list of the variables and their attributes.
- **Enhanced Report** contains the output from the REPORT procedure. This report displays the table and variable attributes. From this report, you can determine the table type, the date on which the table was created and modified, the number of observations, the variable labels, and the variable types. You can sort the rows in the variable table in ascending or descending order by these options:
  - **Variable name**
  - **Variable order in table**
  - **Variable type**
  - **Variable format**
  - **Variable label**

4 Click **Finish** to generate the output.

27

Delete Data Sets and Formats

About the Delete Data Sets and Formats Task

The Delete Data Sets and Formats task enables you to permanently delete data sets, views, and user-defined formats. You can delete multiple items from several different libraries and catalogs at one time.

Note: This task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CATALOG, SQL
Required SAS products	Base SAS
Recommended additional SAS products	None

Delete Data Sets, Views, and User-Defined Formats

1. From the **Server** drop-down list, select the server where the data sets, views, and user-defined formats are stored.
2. In the **Libraries and catalogs** pane, select the library or catalog. The data sets, views, or user-defined formats in the selected library or catalog appear in the list view. By default, these items are sorted by name. Click the column heading to sort the items by the date modified or type.
3. Select one or more of the check boxes for the data sets, views, or user-defined formats that you want to delete. Note

   Note: Below the **Libraries and catalogs** pane, you can see the total number of data sets, views, and user-defined formats that you have selected and the number of selected items in the current folder.

4 Repeat steps 2 and 3 until you have selected all of the items that you want to delete.

5 Click **Run** to permanently delete the selected items.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

28

Discriminant Analysis

About the Discriminant Analysis Task

For a set of observations that contains one or more quantitative variables and a classification variable that defines groups of observations, the Discriminant Analysis task develops a discriminant criterion to classify each observation into one of the groups. The derived discriminant criterion from this data set can be applied to a second data set during the same execution of the discriminant function.

You might want to use this task to classify observations that are found in nature. For example, you could create a SAS data set that contains the species, weight, three different length measurements, height, and width of 159 fish. You could use this data to find a discriminant function based on these six variables that best classifies each fish into one of seven species.

Note: This task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	DISCRIM
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Discriminant Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	performs analyses on each variable in the list. You must assign at least one variable to this role.
Classification variable	defines the groups for the analysis. Classification levels are determined by the formatted values of this variable. You must assign exactly one variable to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for the **Analysis variables** or **Classification variable** roles.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Identifying label	displays the values of the classification variable rather than the observation number. Assigning a variable to this role is effective only when you select the **Resubstitution classification results for every observation** check box. You can assign a maximum of one variable to this role.
Relative weight	uses the values of this variable to weight the corresponding values of the analysis variable. You can assign a maximum of one variable to this role.

Discriminant Analysis: Setting Analysis Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Summary results of cross-validation classification	Classifies each observation in the input data set by using a discriminant function that is computed from the other observations in the data set, excluding the observation being classified. This option applies only when a parametric method is used, which is the default.

Option Name	Description
Display cross-validation classification results for each observation	displays the cross validation classification results for each observation. Note: This option is available only if you select the **Summary results of cross-validation classification** check box.
Resubstitution classification results for every observation	displays the resubstitution classification results for each observation.
Posterior probability error-rate estimates	displays the posterior probability error-rate estimates of the classification criterion based on the classification results.
Univariate test for equality of class means	displays the univariate statistics that are used to test the hypothesis that the class means are equal in the population for each variable.
Multivariate tests for equality of class means	displays the multivariate statistics that are used to test the hypothesis that the class means are equal in the population.
Displays the resubstitution classification results	displays the resubstitution classification results for misclassified observations only. You can specify this option only when the input data set is an ordinary SAS data set.
Displays the cross-validation classification results	displays the cross validation classification results for misclassified observations only.
Performs canonical discriminant analysis	performs a canonical discriminant analysis.
Limits the number of canonical variables to compute	specifies the number of canonical variables to compute. The number that you specify must be less than or equal to the number of variables. If you set this option to 0, then the output includes the canonical correlations but not the canonical coefficients, structures, or means. Let *v* be the number of analysis variables, and let *c* be the number of classification variables. If you do not limit the number of canonical variables to compute, then only min(*v*,*c* - 1) canonical variables are generated. If you include an output data set in your results, v canonical variables are generated. In this case, the last *v* - (*c* - 1) canonical variables have missing values.
Prior probabilities	specifies the method for calculating the prior probabilities. These methods are available: ■ **Equal** specifies that the prior probabilities of group membership should be equal ■ **Proportional sample sizes** specifies that the prior probabilities of group membership should be proportional to the sample sizes ■ **Enter probability for each level** enables you to specify the prior probability for each level of the classification variable. If the sum of the prior probabilities does not equal one, then these probabilities are scaled proportionally to have the sum equal one. Note: The sum of the probabilities that you specify must equal 1. Click **Reset Probabilities** to reset the probabilities to their original values.

Discriminant Analysis: Classifying New Data

Specifying a Data Input Table

In the selection pane, click **Classify New Data** to access these options.

You can select a table (here referred to as the "input table") containing rows that are to be classified according to the discriminant criterion derived from the table that you are analyzing. The input table must contain all of the variables that you assigned to the **Analysis variables** role.

By default, classification results from the input table are included in the output results. The input table is the input data that you used to open the task.

To specify a different input table, select the **New data to be classified** check box and click **Browse**.

If you have assigned a variable to the **Group by variable** role and you want to classify each BY group in the new data by the discriminant function from the corresponding BY group in the original input data, the new data that you select must be sorted in the same order as the input data that you used to open the task.

Note: You need to assign a variable to the **Classification variable** role in order to activate the **New data to be classified** check box.

Note: You must select the **New data to be classified** check box to specify an input table or specify a test variable.

To set the input table options, use these options:

- By default, classification results from the input table are included in the output results. Clear the **Display classification results for the new data** check box to suppress the input table classification results.
- Select `Output table for classification results of the new data` to create a permanent output table that contains all of the data from the input table, plus the posterior probabilities and the class into which each row is classified.
- Select **Output table for group density estimates** to create a permanent output table that contains all of the data from the input table, plus the group-specific density estimates for each row.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

After specifying an input table, you can specify a test variable.

Specifying Test Variable

In the selection pane, click **Classify New Data** to access these options.

After you specify an input table that contains classification data, you can specify a test variable.

To specify a test variable:

1. Select a variable name from the **Classification variable in the new data** drop-down list. This variable is used to determine whether an observation in the input table is misclassified. The variable that you specify here should have the same type (character or numeric) as the variable that is assigned to the **Classification variable** role.

   If you do not select a variable here, the variable assigned to the **Classification variable** role is used.

Note: You might use this option if the variable that is assigned to the **Classification variable** role exists in the input table but has a different name from the original data set. For example, the variable assigned to the Classification variable role is 'Sex', but the classification variable in the input table is 'Gender'.

2 If appropriate, select a variable name from the **Frequency count for test variable** drop-down list. The variable that you select here serves the same function for the test variable as the variable that is assigned to the **Frequency count** role.

3 If appropriate, select a variable name from the **Identifying test label** drop-down list. The variable that you select here serves the same function for the test variable as the variable in the **Identifying label** role. Selecting a variable here is effective only if you select the **Display classification results for the new data** check box.

Discriminant Analysis: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Saving output data	specifies the results to include in an output table. You can choose from these options: ■ The **Posterior probabilities and resubstitution classification** table contains your input data, plus the posterior possibilities and the class into which each observation is classified by resubstitution. ■ The **Posterior probabilities and cross validation classification** table contains your input data, plus the posterior possibilities and the class into which each observation is classified by cross validation. ■ The **Summary statistics and discriminant function** table contains various statistics such as means, standard deviations, and correlations. ■ The **Group specific density estimates for each observation** table contains your input data, plus the group-specific density estimates for each observation. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress reports	Does not display the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

29

Distribution Analysis

About the Distribution Analysis Task

The Distribution Analysis task provides data summarization tools as well as information about the distribution of numeric variables. You can also use it to create a variety of plots, including histograms, probability plots, quantile-quantile plots, and box plots.

You might use this task to create the summary statistics for a product. For example, suppose that you have stored the loan-to-value ratios of 5,840 home mortgages in a SAS data set. Using the Distribution Analysis task, you could create these results:

- a table of summary measures, including moment estimates, and a table of extreme observations
- a histogram that enables you to visualize the distribution of loan-to-value ratios. The histogram reveals features of the distribution, such as its skewness and the peak.
- an analysis of the distribution of the data. This task enables you to run tests for normality and create charts, such as a probability plot.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	UNIVARIATE, BOXPLOT
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Distribution Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

Role Name	Description
Analysis variables	specifies the numeric variables to be analyzed. An analysis variable is required and you can assign multiple variables to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analyses by a variable that you selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight variable. Each value in the variable in this role is used to weight the corresponding values of the variables that you specify as analysis variables. You can assign only one variable to this role.
Classification variables	specifies up to two variables that are used to group the data into classification levels. If you choose to generate a histogram, a probability plot, or a quantile-quantile plot, comparative plots are created. If you choose to generate a box plot, side-by-side box plots are created.

Distribution Analysis: Selecting the Distributions

In the selection pane under the Distributions heading, click **Summary** to access these options.

You can fit several theoretical distributions. After selecting the distribution to compute, select the distribution name in the selection pane to specify the parameter values and line properties for the distribution.

Use the check boxes to select one or more of these estimated density curves:

- **Normal** fits a normal density with a mean (mu) and a standard deviation (sigma).
- **Lognormal** fits a lognormal density with a scale parameter (zeta), a threshold parameter (theta), and a shape parameter (sigma).
- **Exponential** fits an exponential density with a scale parameter (sigma) and a threshold parameter (theta).
- **Weibull** fits a three-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c).
- **Beta** fits a beta density with a scale parameter (sigma), a threshold parameter (theta), and shape parameters (alpha and beta).
- **Gamma** fits a gamma density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (alpha).

- **Kernel** fits nonparametric kernel density estimates.

Select the graphics style to use in the output. You can choose between traditional SAS graphics or the ODS statistical graphics. For more information about ODS graphics, see *SAS Output Delivery System: User's Guide*.

Distribution Analysis: Customizing the Distribution

In the selection pane under the Distributions heading, click the name of the distribution that you want to customize to access these options.

Note: These distribution options are available only if you previously selected the distribution on the **Summary** page or if you select the ***Distribution name*** check box on the page for that distribution.

For the normal, lognormal, exponential, Weibull, beta, and gamma distributions, you can choose to estimate the distribution parameters or to specify them. By default, the task estimates the parameter values for all of the analysis variables.

To specify the parameters for each analysis variable:

1. Clear the **Apply distribution on all variables** check box.
2. Select the variable's check box in the **Analysis variables** box.
3. Specify the parameter values.

Note: To return to the estimated parameter values, click **Use Estimates**.

For nonparametric kernel density estimates, you can specify the bandwidth (c) and the kernel function (k). By default, the task estimates a bandwidth that minimizes the approximate mean integrated square error (MISE). If you do not want to estimate this bandwidth, then clear the **Minimize MISE** check box. To specify the standardized bandwidth parameter for kernel density, enter the bandwidth in the text box. You can specify up to five values with a space between each value. For example, you could enter these values in the text box: `0.5 1.0 1.5`.

Note: If you do not clear the **Minimize MISE** check box, MISE counts as one of the five bandwidth values that you can specify.

You can also specify the kernel function (normal, quadratic, or triangular) that is used to compute the kernel density estimate. The default kernel function is **Normal**.

You can specify kernel functions for up to five estimates. If you specify more than one kernel function, the order of the functions is always normal, quadratic, triangular. To add functions or change the order, you must edit the task code.

If you specify more kernel functions than bandwidths, then the last bandwidth is repeated for the remaining estimates. For example, if you specify 0.5 and 1.0 as the bandwidths and specify normal, quadratic, and triangular as the kernel functions, then the 1.0 bandwidth is repeated for the triangular function.

If you specify more bandwidths than kernel functions, then the last kernel function is repeated for the remaining bandwidths. For example, if you specify 0.5, 1.0, and 1.5 as the bandwidths and specify normal and quadratic as the kernel functions, then the quadratic function is used for the 1.5 bandwidth.

Distribution Analysis: Generating Plots

In the selection pane under the Appearance heading, click **Plots** to access these options.

Note: If a relative weight variable is assigned to an analysis role, you cannot create any plots.

Option Name	Description
Histogram plot	creates a histogram and superimposes density curves for continuous theoretical distributions and for kernel density estimates. To select a distribution, click **Summary** in the selection pane. If you assign classification variables, then a comparative histogram that consists of the component histograms is created. The component histograms enable you to compare the distribution of an analysis variable across the classification levels.
Probability plot	creates a probability plot for a continuous theoretical distribution and superimposes a reference line that corresponds to either the specified or estimated location parameter and scale parameter for theoretical distribution. A probability plot is used to compare the ordered values of a variable to the percentiles of a specified theoretical distribution. If the data distribution matches theoretical distribution, then the points on the plot form a linear pattern. If you assign classification variables, then a comparative probability plot that consists of component probability plots is created. If you do not select a distribution, then the normal distribution is used.
Quantiles plot	creates a quantile-quantile plot (Q-Q plot) for a continuous theoretical distribution and superimposes a reference line that corresponds to either the specified or the estimated location parameter and scale parameter for theoretical distribution. The Q-Q plot is used to compare the ordered values of a variable to the quantiles of the specified theoretical distribution. If the data distribution matches theoretical distribution, then the points on the plot form a linear pattern. If you assign classification variables, then a comparative quantile-quantile plot that consists of component quantile-quantile plots is created. If you do not select a distribution, then the normal distribution is used. **Note:** Q-Q plots are similar to probability plots. Q-Q plots are preferable for graphical estimation of distribution parameters, whereas probability plots are preferable for graphical estimation of percentiles.
Box plot	creates a box plot. If you assign classification variables, then side-by-side box plots are created. A box plot shows the mean, the quartiles, the minimum observation, and the maximum observation for a group of observations.
Text-based plots	creates the following plots: ■ a stem and leaf plot or a bar chart (depending on the number of observations) ■ a box plot ■ a normal probability plot ■ a side-by-side plot if you assigned a variable to the **Group analysis by** role
Axis color, **Background color**, **Axis width**	customizes the appearance of the plot. **Note:** These options are not available for the stem and leaf plot since this is a simple text plot.

Distribution Analysis: Adding an Inset

In the selection pane under the **Appearance** heading, click **Inset** to access these options.

You can enhance a plot by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative histogram, an inset appears in each component histogram.

Distribution Analysis: Specifying Results Options

In the selection pane, click **Tables** to access these options.

By default, the Distribution Analysis report includes the following tables: a basic confidence intervals table, a basic measures table, a tests for location table, and a moments table.

Option Name	Description
Suppress descriptive statistics and capability indices tables	suppresses all descriptive statistics and capability indices tables from the output. Selecting this option does not suppress these tables: ■ tables that are created by the distribution that you selected. You can suppress these tables by selecting the **Suppress distribution tables** check box on the panel for that distribution. ■ ables that are created by the INTERVALS statement. To suppress these tables, you can specify the NOPRINT option. For more information about the INTERVALS statement, see the Help for the CAPABILITY procedure.
Save output statistics to a data set	creates an output data set. The statistics that are included in the final output depend on the tables that are selected in the **Tables** box. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

30

Donut Chart

About the Donut Chart Task

The Donut Chart task creates simple or group charts that show the relative contribution of the parts to the whole. The data appears as wedge-shaped "slices" of a circle. Each slice represents a category of data. The size of a slice represents the contribution of the data to the total chart statistic. For example, a donut chart can show the sales of each store as a fraction of a chain's total sales. A donut chart is similar to a pie chart, except that a donut chart has a hole in the middle of it.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GCHART
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Donut Chart: Selecting a Chart Type

In the selection pane, click **Donut Chart** to access these options.

Option Name	Description
Simple Donut	creates a chart that shows the relative contribution of the parts to the whole. The data appears as wedge-shaped "slices" of a circle. Each slice represents a category of data. The size of a slice represents the contribution of the data to the total chart statistic.
Group Donut	produces a separate donut chart for each unique value of the variable that is assigned to the **Group donuts by** role.

Donut Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Note: The available roles depend on the chart type that you selected.

Role Name	Description
Column to chart	The values of the column that you assign to this role determine the different slices in donut charts. The column can be either character or numeric.
Group donuts by	A donut chart is created for each unique value of the column that you assign to this role. All the charts are displayed on the same page.
Sum of	The column that you assign to this role determines the sizes of the sections in the donut chart. If you do not assign a column to this role, then the frequency of each value of the **Column to chart** column is used to determine the sizes of the sections.
Group analysis by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Each donut chart is displayed on a new page.

Donut Chart: Setting Appearance Options

Setting Donut Chart Options

In the selection pane under the Appearance heading, click **Donuts** to access these options.

Role Name	Description
Specifying colors for the slices	specifies the color of the slices. You can choose from these options: ■ **Default color scheme** uses the default color scheme for the slices. The default color scheme is determined by the current style. You can specify the style by using the Style Manager or the Properties dialog box for the result. In SAS Enterprise Guide, you can also specify the style for each type of result from the Options dialog box. In the SAS Add-In for Microsoft Office, you can apply styles to the SAS Report and HTML formats by using the **SAS Add-In Options** dialog box. ■ **Existing colors (Specified in a previous version)** enables you to use colors that were specified in a previous version of SAS Enterprise Guide or the **SAS Add-In for Microsoft Office**. This option is available only for SAS content that you have migrated from a previous release. ■ **Custom colors** enables you to specify the custom colors to use for the slices. Using the drop-down lists, you can specify up to 12 colors for the bar chart. If your chart needs to use more than 12 colors, then the additional colors are selected from the current default color scheme. Click **Reset** to reset the custom colors back to the default. **Note:** Depending on the type of donut chart, not all of the custom colors might be used.
Determining the number of slices	specifies the number of slices in the chart. By default, the number of slices is determined automatically. For continuous columns, the values are divided into ranges, with one slice for each range. Slices that would represent 4% or less of the total are combined into a slice named “Other.” To specify how many slices appear in the chart, select the **Specify number of slices** check box.

Specifying the Layout

In the selection pane under the Appearance heading, click **Layout** to access these options.

Select whether to use the name, the percentage, the statistic value, or a combination of the three in the label for each slice. For each component of the label, select one of these positions for the label:

- inside the slice
- outside the pie or donut
- outside the slice with an arrow that points to it

You can choose to make the font color for the slice labels the same color as the slices. If you want the labels to appear inside the slices, then do not request this option.

You can specify the percentage and the label for the "Other" slice. This slice is a collection of all the midpoints that have chart statistic values less than or equal to the percent-of-total. The default value of the percent-of-total is 4. Therefore, any slice that represents 4% or less of the total is put in the "Other" slice.

You can specify a text string up to 16 characters long as the label of the "Other" slice. The default label is Other. The "Other" slice is the last slice in the pie, regardless of the order of the slices.

If only one midpoint falls into the "Other" category, then the slice for that midpoint remains in its original position in the pie and retains its original label. For example, suppose a pie that shows the percent sales of beverages has these slices and percent values: Coffee 35%, Tea 15%, Soda 5%, and Milk 45%. If the percent-of-total is 5, the slice for soda remains the third slice in the pie instead of becoming the last slice, and this slice is labeled "Soda" instead of "Other."

For grouped charts, use the **Grouped Layout** area to arrange the charts. Specify the number of columns across the output area or the number of rows down the output area.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. Note: If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. Note: This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. Note: The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Setting Advanced Options

In the selection pane under the Appearance heading, click **Advanced** to access these options.

Select from the drop-down list the statistic to use when calculating the slices in the chart. If you do not assign a variable to the **Sum of** role, you can choose to calculate the frequency or percentage of each slice. By default, the frequency is calculated. If you do assign a variable to the **Sum of** role, then you can choose to calculate the sum or average of each slice. By default, the sum is calculated.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

31

Download Data Files to PC

About the Download Data Files to PC Task

The Download Data Files to PC task enables you to select one or more SAS data sets (*.sas7bdat) from one or more SAS servers and copy them to your local Microsoft Windows file system.

The task generates a log that summarizes the data files that were copied, the byte size, timings for each file and for all of the files as a group, and any errors that occurred during the transfer.

Download Data Files to PC: Select the SAS Data Sets to Download

1 Click **Add** to select the SAS data sets that you want to copy. The Open dialog box appears.

2 Navigate to the appropriate server and library, and select the name of the file that you want to copy. To select more than one file, hold down the Ctrl key while selecting the files that you want to add.

3 Click **Open** to add the files to the list of data sets that you want to copy.

   Note: You can select data files from multiple SAS servers by clicking **Add** again and selecting data from another server.

   Note: All files that are copied are given lowercase names in the target folder.

4 Click **Next** to specify the download options.

Download Data Files to PC: Specify the Download Options

1 Before you can specify the download options, you must select the data sets that you want to download.

2 Enter the path of the folder to which you want to copy the files. The default location is your My Documents folder.

3 To overwrite any existing files on your local computer that have the same name as the files that you are copying, select the **Overwrite existing files with the same name** check box. By default, existing files are not overwritten.

4 Click **Finish** to copy the data sets.

Note: If the source library is a Base SAS engine library with a single file path, then the file transfer will go directly from that library path to the target folder. If the source library is not a Base SAS engine library or it is a concatenated library, then the file transfer goes first to a staging area in WORK, and then it is copied to the target folder by using PROC SQL.

32

Factor Analysis

About the Factor Analysis Task

Factor analysis performs a variety of common factor and component analyses and rotations. Input can be multivariate data, a correlation matrix, a covariance matrix, a factor pattern, or a matrix of scoring coefficients.

You might want to use this task to perform a common factor analysis on an annual employee review. For example, suppose that 103 police officers were rated by their supervisors on 14 scales (variables). You can conduct a common factor analysis on these variables to see what latent factors are operating behind these ratings.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	FACTOR
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	None

Factor Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the variables to be analyzed. You must assign at least one variable to this role.
Partial variable	specifies the variable that is used to partial out the variables to be analyzed. This bases the analysis on a partial correlation or covariance matrix.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group the table by a variable that you have already selected for **Analysis variables**.
Frequency count	specifies one variable that represents the frequency of occurrence for the other values within the observation. This causes the task to treat each row as if it occurs n times, where n is the value of the frequency count for that observation. The total number of observations is considered to be equal to the sum of the variable when the degrees of freedom for significance probabilities are determined. If the value of the frequency variable is missing or less than one, the observation is not used in the analysis. If the value is not an integer, only the integer portion is used.
Relative weight	specifies one variable to use as relative weights for each observation in the input data set. This is often done when the variance that is associated with each observation is different and the values of the relative weight variable are proportional to the reciprocals of the variances. If a relative weight value is negative or is missing, it is excluded from the analysis.

Factor Analysis: Setting Factoring Method Options

In the selection pane, click **Factoring Method** to access these options.

Option Name	Description
Factoring method	specifies the factoring method. You can choose from these options: ■ Principal component analysis performs principal component analysis if all priors equal one. Otherwise, a principal factor analysis is performed. This method is the default, except for data sets of type FACTOR. ■ **Harris component analysis** performs Harris component analysis of S-1RS-1, a noniterative approximation to canonical component analysis. ■ **Image covariance matrix** performs principal component analysis of the image covariance matrix. A nonsingular correlation matrix is required. ■ **Alpha factor analysis** performs alpha factor analysis. ■ **Maximum-likelihood factor analysis** performs maximum likelihood factor analysis with an algorithm due, except for minor details, to Fuller. A nonsingular correlation matrix is required. ■ **Iterated principal factor analysis** performs iterated principal factor analysis. ■ **Unweighted least squares factor analysis** performs unweighted least squares factor analysis. ■ **Existing factor pattern** reads a factor pattern from a data set type of FACTOR, CORR, UCORR, COV, or UCOV. This method is the default for data sets of type FACTOR. ■ **Factor scoring coefficients** reads the scoring coefficients from a data set type of FACTOR, CORR, UCORR, COV, or UCOV. The data set must also contain either a correlation or a covariance matrix. After you select a method, you can set the **Numerical properties** and **Number of factors** options in the table. If you change two or more of the **Number of factors** options, the number of factors retained is the minimum number satisfying any of the following criteria: ■ Smallest eigenvalue for which a factor is retained ■ Number of factors to retain ■ Percent of common variance accounted for by retaining factor
Matrix options	specifies the matrix options ■ **Factor covariance matrix instead of correlation matrix** uses the covariance matrix instead of the correlation matrix for factoring when the factoring method is principal component analysis, image covariance matrix, iterated principal factor analysis, or unweighted least squares factor analysis. ■ **Factor a weighted matrix** uses a weighted correlation or covariance matrix for factoring when the factoring method is principal component analysis, image covariance matrix, iterated principal factor analysis, or unweighted least squares analysis. **Note:** This option is available only if the factoring method is principal component analysis, image covariance matrix, iterated principal factor analysis, or unweighted least squares factor analysis and the input data set is of type CORR, UCORR, COV, UCOV, or FACTOR. ■ **Correct the resulting matrix for the mean** uses the intercept from the analysis; covariances or correlations are corrected for the mean. By default, this option is selected.

Option Name	Description
Divisor for variance	specifies the divisor that is used in the calculation of variances and covariances. By default, the divisor for the variance is the degrees of freedom.

Factor Analysis: Setting Communality Options

In the selection pane, click **Communality** to access these options.

Option Name	Description
Prior communality estimates	Select one of these options: ■ **Proportional to squared multiple correlations** sets the prior communality estimates proportional to the squared multiple correlations but adjusts them so that their sum is equal to that of the maximum absolute correlations. ■ **Squared multiple correlation with all other columns** sets the prior communality estimate for each variable to its squared multiple correlation with all other variables. ■ **Maximum absolute correlation with other columns** sets the prior communality estimate for each variable to its maximum absolute correlation with any other variable. ■ **Set all priors to one** sets all prior communalities to 1.0. ■ **Input from PRIORS or COMMUNAL row in input data table** reads the prior communality estimates from the first observation when the _TYPE_ variable contains PRIORS or COMMUNAL and the data set type is FACTOR. Note: This option is available only if the factoring method that you selected is principal component analysis, alpha factor analysis, maximum likelihood factor analysis, iterated principal factor analysis, or unweighted least squares factor analysis and the input data set is of type FACTOR. ■ **Set priors to a pseudo-random between 0 and 1** sets the prior communality estimates to pseudo-random numbers that are uniformly distributed between 0 and 1. ■ **Explicitly specify priors values for each column** enables you to select the prior communality estimate for each variable from the input box and then to specify a new value. By default, the prior communality estimate for each variable is `0`. Note: The default prior communality estimates depend on the factoring method that you use.

Option Name	Description
Heywood options	By default, the factoring methods alpha factor analysis, maximum likelihood factor analysis, iterated principal factor analysis, and unweighted least squares factor analysis stop iterating and set the number of factors to 0 if an estimated communality exceeds 1. You can also choose from these options: ■ **Set any communality that is greater than one to one** sets to 1 any communality greater than 1, allowing iterations to proceed. ■ **Allow communalities to exceed one** allows iterations to proceed when communalities exceed 1. Convergence problems are possible because communalities can become extremely large, and ill-conditioned Hessians might occur.

Factor Analysis: Setting Rotation and Plot Options

In the selection pane, click **Rotation and Plots** to access these options.

Option Name	Description
Rotation method	specifies the rotation method. You can choose from these methods. Depending on the method that you select, additional options might be available. ■ **No rotation** specifies that no rotation be performed, leaving the original orthogonal solution. This is the default method. ■ **Orthogonal equamax** specifies orthogonal equamax rotation. This corresponds to using general orthomax with an orthomax weight equal to *n*/2, where *n* is the number of factors. ■ **Harris-Kaiser case II orthoblique** specifies Harris-Kaiser case II orthoblique rotation. You can specify the power of the square roots of the eigenvalues that are used to rescale the eigenvectors, assuming that the factors are extracted by the principal factor method. For other extraction methods, the eigenvectors are replaced by the normalized columns of the unrotated factor matrix, and the eigenvalues are replaced by the column-normalizing constants. The default eigenvalue of zero yields the independent cluster solution, in which each variable tends to have a large loading on only one factor. ■ **General orthomax** specifies the orthomax rotation with a default orthomax weight equal to zero. There is no restriction on valid values for the orthomax weight, although the most common values are between zero and the number of variables. An orthomax weight equal to one results in the varimax rotation. ■ **Orthogonal Parsimax** specifies orthogonal parsimax rotation. This corresponds to using general orthomax with an orthomax weight equal to [(*nvar* ×(*n* - 1))/(*nvar* + *n* - 2)], where *nvar* is the number of variables and *n* is the number of factors. ■ **Oblique Procrustes** specifies oblique Procrustes rotation with the target pattern that is provided by the TARGET data set. The unrestricted least squares method is used, with factors scaled to unit variance after rotation. ■ **Oblique promax** specifies oblique promax rotation. You can set the desirable pre-rotation method to orthogonal or oblique. You can also change the power that is used to compute the target pattern from the default value of `3`. ■ **Orthogonal quartimax** specifies orthogonal quartimax rotation, which corresponds to using general orthomax with an orthomax weight equal to zero. ■ **Orthogonal varimax** specifies orthogonal varimax rotation, which corresponds to using general orthomax with an orthomax weight equal to one.
Method for normalizing rows of the factor pattern	specifies the method for normalizing the rows. You can choose from these options: ■ **Kaiser's normalization** uses Kaiser's normalization which is the default method. ■ **Weight by Cureton-Mulaik technique** weights the rows by using the Cureton-Mulaik technique. ■ **Rescale to represent covariances** rescales the rows of the pattern matrix to represent covariances instead of correlations. ■ **Do not normalize** prevents normalization from occurring.

Option Name	Description
Plots to show	specifies the plots to include in the results. You can choose from these options: ■ **Plot the factor pattern (after rotation)** displays a plot of the factor pattern after rotation. Use the **Number of factors to plot** option to set the number, *n*, of factors to be plotted so that all pairs of the first *n* factors are plotted, producing a total of *n*(*n* - 1)/2 plots. Enter an integer value directly or use the up and down arrows to adjust the value by one. The default value is 2. ■ Plot the unrotated factor pattern displays a plot of the unrotated factor pattern. ■ Show a scree plot of the eigenvalues displays a scree plot of the eigenvalues. By default, no plots are shown. After you select the plots to show, you can also specify the number of factors to plot. You can choose from these options: ■ **Use all factors** plots all the factors. This is the default. ■ **Number of factors to plot** enables you to specify how many factors to plot (*n*). The smallest value that you can specify is 2. When you specify a value, all pairs of the first n factors are plotted, producing a total of *n*(*n* - 1)/2 plots.

Factor Analysis: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Save output data	Select the **Factors** or **Statistics** check boxes for the output data that you want to save permanently in the **Save output data** area. If you select **Statistics**, you can choose to exclude the correlation matrix from the output data set. The correlation matrix is excluded only if the correlation matrix is not required for other requested output (such as if the scores or the residual correlations are displayed) and if you select **Existing factor pattern** as the factoring method. This method is the default for data sets of type FACTOR. Note: The **Factors** option is not available if the input data set is of type CORR, UCORR, COV, UCOV, or FACTOR. This option is also not available if you assigned a variable to the **Partial variables** role. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Option Name	Description
Display output	Select one or more of the following options in the **Display output** area: ■ **Residual correlation matrix** displays the residual correlation matrix and the associated partial correlation matrix. The diagonal elements of the residual correlation matrix are the unique variances. ■ **Eigenvectors** displays the eigenvectors of the reduced correlation matrix, in which the diagonal elements are replaced with the communality estimates. When you use the maximum likelihood factor analysis factoring method, the eigenvectors are for the weighted reduced correlation matrix. ■ **Factor scoring coefficients** displays the factor scoring coefficients. The squared multiple correlation of each factor with the variables is also displayed except in the case of unrotated principal components. ■ **Reorder matrix rows by highest absolute loading** reorders the rows (variables) of various factor matrices in the output. Variables with their highest absolute loading (reference structure loading for oblique rotations) on the first factor are displayed first, from largest to smallest loading, followed by variables with their highest absolute loading on the second factor, and so on. ■ **Means and standard deviations of input columns** displays means, standard deviations, and the number of observations. ■ **Correlation matrix of input columns** displays the correlation matrix or partial correlation matrix. ■ **Kaiser's measure of sampling adequacy** produces the partial correlations between each pair of variables, controlling for all other variables (the negative anti-image correlations), and Kaiser's measure of sampling adequacy. ■ **Input factor pattern** displays the input factor pattern and related statistics. In oblique cases, the reference and factor structures are computed and displayed. **Note:** This option is available only if you selected the existing factor pattern method. ■ **Scoring coefficients** displays the scoring coefficients and related statistics. In oblique cases, the reference and factor structures are computed and displayed. **Note:** This option is available only if you selected the factor scoring coefficients method. By default, no factor result or related statistics computations are selected.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

33

Forecast Studio Create Project

About the Forecast Studio Create Project Task

The Forecast Studio Create Project wizard helps you specify the forecasting variables, choose whether to forecast your data hierarchically, and specify the forecast horizon for a new SAS Forecast Studio project. **SAS Forecast Studio** is a forecasting application that is designed to speed the forecasting process through automation. The software provides for the automatic selection of time series models that are used in forecasting time-stamped data. To save this project, you must have access to SAS Forecast Server.

Note: To create a forecasting project that uses SAS data from your PC, you should select the data source through the file system rather than selecting a data source on your local computer.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	SAS Forecast Server
Recommended additional SAS products	None

Forecast Studio Create Project: Verify the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to select the forecasting variables.

Forecast Studio Create Project: Select the Forecasting Variables

1 Before you can select the forecasting variables, you must verify the data.

2 Assign variables to these roles:

Time ID
This variable specifies the time ID for the data. The time ID variable is a variable in the input data set that contains the SAS date, datetime, or time value of each observation. This variable is used to determine the frequency and ordering of the data and to extrapolate the time ID values for the forecasts. You can assign only one variable to this role, and it must be either a date variable, a datetime variable, or a numeric variable that contains date or datetime values.

Forecast (Dependent variables)
The columns that you assign to this role specify which variables to model and forecast. The Forecast variables are also called dependent variables. You must assign at least one numeric variable to this role. Use the up and down arrows to specify the order of the variables.

3 (Optional) Assign one or more numeric variables to the **Predict using (Independent variables)** role. The column that you assign to this role specifies an explanatory, input, predictor, or causal factor variable. The Predict using variables are also called independent variables. Specify the order of the variables.

4 Click **Next** to select the classification variables.

Forecast Studio Create Project: Select the Classification Variables

1 Before you can select the classification variables, you must select the forecasting variable.

2 Assign variables to the **For each value of (Classification variable)** role. Variables that are assigned to this role are called classification variables. Observations that have the same value for the classification variable are grouped. Assigning a classification variable enables you to obtain separate analyses for groups of observations. You can assign character and numeric variables to this role.

The order of the classification variables describes the structure of the hierarchy. Use the up and down arrows to specify the order of the variables.

3 Specify how to display the output. You can choose from these options:

- **List view** does not create a hierarchy of the data. The forecast series are organized in a list.
- **Hierarchy** uses the values of the classification variables to organize the data into a hierarchy. This option is selected by default.

Note: If you assign multiple variables to the **Forecast (dependent variables)** role, the **Hierarchy** option is not available.

4 If you choose to display the output as a hierarchy, then you can specify the hierarchy settings by clicking **Settings**. The Hierarchy Settings window appears.

You can choose from these options:

Aggregation method

Specify the statistic for aggregating data across the levels of the hierarchy. By default, all the dependent and independent variables are listed in the table. In the **Aggregation** column, select the aggregation statistic to use for each variable.

For example, your data set contains the sales for a group of products. You might want to know the total sales for a category. In this case, you would choose **Sum** as the aggregation statistic. If your data contains the price of each product, you might want to know the average price for a product line. In this case, you would choose **Average** as the aggregation statistic.

Reconciliation level

Select the level in the hierarchy that contains the data that you want to forecast. You can choose from these options:

- **Bottom Up** uses the data at the lowest level of the hierarchy to generate the forecasts. These forecasts are then used to generate the reconciled forecasts for the higher levels in the hierarchy.

  The Bottom Up method enables you to see any patterns (such as seasonality) in the data. However, because you are at the lowest level of the hierarchy, you can also have too much noise or randomness in the data. Also, these forecasts might fail because the data at the lowest level of the hierarchy can be sporadic or too sparse.

- **Middle Out** adds up the data from the lower levels and then uses these values to generate the forecasts for the middle level. Because some hierarchies have more than one middle level, you need to specify the level that you want to use. The forecasts at the middle level are used to generate forecasts for both the higher and lower levels.

- **Top Down** adds up the data from the lowest levels in the forecast and then uses this value to generate the forecasts. The forecast at the highest level is used to generate the forecasts for lower levels in the hierarchy.

  The Top Down method enables you to remove the excessive noise from the data at the lower levels of the hierarchy. However, you also might lose the pattern (such as the seasonality) in the forecast.

Disaggregation method

Select the disaggregation method for the hierarchy. You can choose from these options:

- **Forecast Proportions** reconciles forecast based on the historical mean. This is the default.
- **Equal Split of the Difference** reconciles forecast based on forecast deviations between levels.

5 Specify the number of periods into the future for which multi-step forecasts are made in the **Number of periods to forecast (horizon)** box. The default value is `12`.

6 Click **Next** to save the project.

Forecast Studio Create Project: Save the Project

1 Before you can save the project, you must select the classification variable.

2 Enter a name for the project in the **Save as** box. The project name must be a valid SAS name. The project name can be 32 characters long, and it must start with a letter (A–Z). Subsequent characters can be letters or numeric digits (0–9). Both upper and lowercase letters are valid. For more information about SAS naming conventions, see *SAS Language Reference: Concepts*

3 Click **Finish** to create the SAS Forecast Studio project. After this project is created, you can open it in SAS Forecast Studio.

34

Forecast Studio Open Project

About the Forecast Studio Open Project Task

The Forecast Studio Open Project wizard helps you open the selected series from an existing SAS Forecast Studio project and specify how to display the results. SAS Forecast Studio is a forecasting application that is designed to speed the forecasting process through automation. The software provides for the automatic selection of time series models that are used in forecasting time-stamped data.

With this task, you can open a project that was created in SAS Forecast Studio and perform additional analyses. To open a forecasting project, you must have access to SAS Forecast Server.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	SAS Forecast Server
Recommended additional SAS products	None

Forecast Studio Open Project: Select the Forecasting Project

In the **Forecast projects** pane, you can see a list of all the available projects and the date on which the project was last modified. Select the forecasting project that you want to open. You can open only one project at a time.

The details for a project are not displayed by default. To see the description, dependent variables, and classification (or BY) variables that are associated with the selected project, click **Details**. Because this

information is stored on the forecasting server, it might take a few seconds for the details to appear. To hide the details, click **Hide**.

Click **Next** to select the forecasting series.

Forecast Studio Open Project: Select the Forecasting Series

1 Before you can select the forecasting series, you must select the forecasting project.

2 In the **Select series to display** pane, select the check box for the forecast series to open.

   The series are presented in a list if the project data was not forecast hierarchically. The series are presented in a hierarchy if the data was forecast hierarchically. Click **Expand All** to expand all of the nodes in the hierarchy. To collapse the nodes, click **Collapse All**.

3 Click **Next** to select the result types.

Forecast Studio Open Project: Select the Result Types

1 Before you can select the result types, you must select the forecasting series..

2 Select the result types to open. You can choose from these options:

   - **Forecast Plot** displays the historical values, the fitted values, and the reconciled forecast for the selected model.
   - **Forecast Table** displays the model for the statistical forecast, the reconciliation adjustment, the reconciled forecast, the manual override values, and the final forecast.
   - **Model Selection Table** shows all of the models that have been defined for the current project.
   - **Model Parameter Estimate Table** displays estimates, standard errors, and significance tests for each of the model parameters.
   - **Component Table** displays tables for each of the components that are used by the model.
   - **Statistics of Fit Table** displays all available statistics of fit for the forecasting model.
   - **Forecast Summary Table** displays forecast values of the dependent variable for the time range of the forecast horizon.

3 Click **Next** to specify titles and footnotes.

Forecast Studio Open Project: Specify Titles and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

35

Forecast Studio Override Project

About the Forecast Studio Override Project Task

SAS Forecast Studio is a forecasting application that is designed to speed the forecasting process through automation. The software provides for the automatic selection of time series models that are used in forecasting time-stamped data. To open a forecasting project in SAS Enterprise Guide or the SAS Add-In to Microsoft Office, you must have access to SAS Forecast Server.

The Forecast Studio Override Project wizard enables you to submit overrides for the forecast data in an existing SAS Forecast Studio project. An override is a replacement value that you specify in place of a forecasted value. You can submit overrides for future values only.

After you submit an override, SAS Forecast Server reconciles the project. How the override is applied depends on whether the data in the project is forecasted hierarchically.

- If the data in the project is forecasted hierarchically, then the override is an adjustment that is done with respect to the reconciled statistical forecast.
- If the data in the project is not forecasted hierarchically, then the override is an adjustment that is done with respect to the statistical forecast.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	SAS Add-In for Microsoft Office, SAS Forecast Server
Recommended additional SAS products	None

Forecast Studio Override Project: Select the Forecasting Project

In the **Forecast projects** pane, you can see a list of all the available projects and the date on which the project was last modified. Select the forecasting project that you want to open. You can open only one project at a time.

The details for a project are not displayed by default. To see the description, dependent variables, and classification (or BY) variables that are associated with the selected project, click **Details**. Because this information is stored on the forecasting server, it might take a few seconds for the details to appear. To hide the details, click **Hide**.

Click **Next** to select the forecasting series.

Forecast Studio Override Project: Select the Forecasting Series and Submit the Overrides

Select the Forecasting Series

1 Before you can select the forecasting series, you must select the forecasting project.

2 In the tree view, select the series for which you want to submit an override.

   If the project data was forecast hierarchically, the series are presented in a hierarchy. Click **Expand All** to expand all of the nodes in the hierarchy. To collapse the nodes, click **Collapse All**.

   The series are presented in a list if the project data was not forecast hierarchically.

3 After you have selected the series, click **Override Forecast Data** to create and submit the overrides.

4 Repeat steps 2 and 3 for all the series for which you want to submit an override.

5 After you have created and submitted your overrides, click **Finish** to close this wizard.

Create and Submit the Overrides

Two graphs are at the top of the Override Forecast Data dialog box. In the graph on the left, the blue line shows the forecast and the circles represent the historical data. In the graph on the right, the dark blue line shows the forecast; the lighter blue lines are the confidence intervals for the forecast; and the black plus signs represent any overrides that you specify.

To create and submit the overrides:

1 In the **Edit Overrides** table, select the time period where you want to add or edit an override. You can select multiple time periods by pressing Ctrl and then selecting the rows in the table. 2. 3. A Note

2 To create an override for an individual time period, click the cell in the **Overrides** column for that time period. Enter the override value and press Enter.

   To calculate overrides for the selected time periods, click **Calculate Overrides**. The Override Calculator dialog box appears.

To remove the overrides that you specified during this session, click **Reset**. Any overrides that you specified previously or that were specified in SAS Forecast Studio are not removed.

3 After you create overrides for the current series, click **Submit**. SAS Forecast Server reconciles the project. After the project is reconciled, the Override Forecast Data window closes, and you return to the Select the forecasting series step in the wizard. In the tree view, the Override icon appears next to the series for which you submitted the overrides.

Note: If SAS Forecast Server detects an override conflict when reconciling the project, then an error message appears, stating that the project cannot be reconciled and that you must use SAS Forecast Studio to reconcile the conflicts.

Calculating Overrides for the Selected Time Periods

You can calculate the overrides for one or more time periods by using these options:

Option Name	Description
Set relative to the forecast	enables you to base the override values on the specified increase or decrease of the reconciled statistical forecast. **Note:** If you did not forecast your data hierarchically or there is no reconciliation forecast available, then you can select this option to base the override values on the specified increase or decrease of the statistical forecast. Using the drop-down lists and text box, specify the number of units or percentage to increase or decrease the current value. Select the **Increase** or **Decrease** option. In the text box for the option that you selected, specify the value, and in the drop-down list, specify whether this value is in units or a percent. For example, you want to create an override that is 10% greater than the reconciled statistical forecast. Select the Increase option. In the text box, type 10, and from the drop-down list, select **% (percent)**. The override values are 10% greater than the reconciled statistical forecast.
Distribute a value among selected periods	distributes a single override value across all the selected periods. Specify this override value in the text box. How the override value is applied depends on the option that you choose. You can choose from these options: ■ **Distribute proportional to current forecast for each period** - the value that you specify is divided proportionally between the selected time periods. For example, you select two time periods with forecasts of 550 and 430. You specify a value of 100 and you choose to divide this value proportionally between the two time periods. Now, the forecasts are 606.12 and 473.88 ■ **Distribute equally among period**s - the value that you specify is divided equally among the selected time periods. For example, you select two time periods with forecasts of 550 and 430. You specify a value of 100 and you choose to divide this value equally between the two time periods. Now, the forecasts are 600 and 480.
Assign a value to each selected period	applies a single override value to all the selected periods. Specify this override value in the text box.

After you have calculated the overrides for the selected time periods, click **OK** to submit the overrides.

Resolve Override Conflicts

If SAS Forecast Server detects an override conflict when reconciling the project, then an error message appears, stating that the project cannot be reconciled and that you must use SAS Forecast Studio to reconcile the conflicts. In SAS Forecast Studio, you can see all of the overrides that were submitted, and you can choose which overrides to apply to the forecasting data.

To resolve override conflicts:

1. Click **Cancel** to exit the Forecast Studio Submit Overrides wizard and close Microsoft Excel.
2. Open the project in SAS Forecast Studio. When you open the project in SAS Forecast Studio, the "Reconciliation is out of date" warning appears at the top of the workspace.
3. .Next to this warning, click **Update**. The Override Conflicts dialog box appears.
4. To resolve the override conflicts, you can modify the overrides by using the Overrides Conflicts dialog box, or you can close this dialog box and update overrides by using the data table.

For more information about resolving override conflicts, see the Help for SAS Forecast Studio.

36

Generalized Linear Models

About the Generalized Linear Models Task

The Generalized Linear Models task is an extension of traditional linear models that enables you to model data that is not normally distributed, such as counts or measured proportions that are considered discrete. You can also use the generalized linear model task to model data for which the mean has been restricted to a range of values or data for which the variance is nonconstant for all observations.

You might want to use this task to perform these tasks:

- compare the life span of products from two different manufacturers. For example, suppose that you have the failure times of machine parts, some of which are manufactured by manufacturer A and some by manufacturer B. You can use this task to test for differences between the two manufacturers in the life span of their machine parts.
- fit a model to data that is measured on an ordinal scale. For example, suppose that you have a data set that contains the results of a hypothetical taste test of three brands of ice cream. The three brands are rated for taste on a five-point scale, from very good (vg) to very bad (vb). An analysis is performed to assess the differences in the ratings for the three brands.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GENMOD
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Generalized Linear Models: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variables	specifies the variables to use as the dependent (response) variables. You must assign at least one variable to this role.
Quantitative variable	specifies the variables to use as the continuous independent (explanatory) variables (also known as effects).
Classification variables	specifies the variables to use as the discrete independent effects. Variables that you assign to this role can be numeric or character, but they should have a limited number of discrete values.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies one variable to use as relative weights for the corresponding values of the variable that you selected as the response variable. You can assign a maximum of one variable to this role.

Role Name	Description
Relative offset	specifies one variable to use as the offset variable. **Note:** You can assign a maximum of one variable to this role. Note You cannot offset by a variable that you have already selected as a dependent variable, an independent variable, or a classification variable.

Generalized Linear Models: Building a Model

In the selection pane, click **Model** to access these options.

By default, no effects are specified, which results in the task fitting an intercept-only model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects.

Specify Main Effects

1 Select the variable name in the **Class and quantitative variables** box.

2 Click **Main** to add the variable to the **Effects** box.

Specify Crossed Effects (Interactions)

1 Select one or more variable names. To select more than one variable, press Ctrl while you select the additional variable names.

2 Click **Cross**. If you select only one variable, then the variable is crossed with itself.

Specify Nested Effects

1 Select one or more main effects or crossed effects that you want to nest.

2 Select one or more Class variables within which you want to nest the main effects or crossed effects.

3 Click **Nest**. The nested effect appears in the **Effects** box in the form *variable 1*(*variable 2*).

Create Multiple Crossings of Two or More Selected Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Factorial** to add the factorial effects to the **Effects** box. By default, each combination of two variables is used.

Create Polynomial Crossing of Quantitative Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Polynomial** to add the polynomial effects to the **Effects** box.

Remove the Effects of the Intercept from the Analysis

Clear the **Include intercept check** box. If you clear this check box, then you must have specified at least one effect. By default, the intercept is included.

Generalized Linear Models: Setting Model Options

In the selection pane, click **Model Options** to access these options.

Option Name	Description
Distribution	specifies the probability distribution of the response variable in your model. Note: Binomial distributions are available only if the dependent variable has two levels. Note: f a character variable is assigned to the **Dependent variables** role, only the multinomial and binomial distributions are available.

Option Name	Description
Link functions	describes how the expected value is related to the linear predictor. Several link functions are available, depending on the model that you are constructing. From the **Link function** drop-down list, you can choose from these options: ■ **None** ■ **Identiy** ■ **Logit** ■ **Probit** ■ **Log** ■ **Complementary log-log** ■ **Power with number** Note: The **Cumulative logit**, **Cumulative complementary log-log**, and **Cumulative probit** link functions are available only when you select a multinomial distribution. If you do not select a link function from the **Link function** drop-down list and there is a user-defined link function, then the user-defined link function is used. If you did not specify a user-defined link function, then the default link function for the model distribution is used. ■ For a binomial distribution, the default link function is logit. ■ For a gamma distribution, the default link function is inverse (power(-1)). ■ For an inverse Gaussian distribution, the default link function is inverse squared (power(-2)). ■ For a multinomial distribution, the default link function is cumulative logit. ■ For a negative binomial distribution, the default link function is log. ■ For a normal distribution, the default link function is identity. ■ For a Poisson distribution, the default link function is log. Note: If you select **None** for the distribution and for the link function, then the normal distribution is used with the identity link function in the analysis.
Effects tests	specifies the statistics to include in the analysis. ■ Select the **Compute a Type 1 or sequential analysis** check box to show the likelihood ratio statistic between successive pairs of models. The sequentially fitted models begin with the null (intercept term only) model and continue up to the models that you defined. ■ Select the **Compute statistics for Type 3 contrasts** check box to show the contrast statistics for each effect in the models that you defined. The default analysis is to compute the likelihood ratio statistics for the contrasts. Wald statistics are also computed if the **Compute Wald statistics** check box is selected.
Confidence limits	specifies the confidence level for the predicted values. The default value is a 95% confidence level.
Estimate details	specifies whether to display the parameter estimate correlation matrix, the parameter estimate covariance matrix, or both.

Option Name	Description
Confidence intervals	specifies whether to use a two-sided confidence interval. ■ Select the **Two-sided** check box to compute two-sided confidence intervals for all model parameters based on the profile likelihood function (also known as the partially maximized likelihood function). ■ Select the **Two-sided Wald** check box to compute two-sided Wald confidence intervals for all model parameters based on the asymptotic normality of parameter estimates.

Generalized Linear Models: Selecting Post Hoc Tests

In the selection pane, click **Post Hoc Tests** to access these options.

For each effect in the **Effects to estimate** box, specify options in the **Options for means tests** box.

Generalized Linear Models: Generating Plots

In the selection pane, click **Plots** to access these options.

By default, all appropriate plots for the current data selection are included in the output. However, you can choose which plots to include in the output by selecting the **Custom lists of plots** option. You can choose from the following options:

Option Name	Description
Cook's D Plot	creates a plot of Cook's D statistic versus the observation number.
DFBETA plots	creates plots of the difference in the parameter estimate due to deleting the observation.
DFBETAS plots	creates plots of the standardized difference in the parameter estimate due to deleting the observation. It is useful to assess the effect of an individual observation on each estimated parameter of the fitted model.
Leverage plot	creates a plot of the value of the corresponding diagonal element in the H matrix. The diagonal elements of the matrix are useful in detecting extreme points in the design space where they tend to have larger values.
Predicted value plot	creates a plot of the predicted values.
Pearson residual plot	creates a plot of the Pearson (chi) residual that is used to identify observations that are poorly accounted for by the model versus the predicted values.
Deviance residual plot	creates a plot of the deviance residual that is used to identify poorly fitted values versus the predicted values.

Option Name	Description
Likelihood residual plot	creates a plot of the likelihood residual that is used to identify poorly fitted values versus the predicted values.
Raw residual plot	creates a plot of the raw residuals versus the predicted values.
Standardized Pearson residual	creates a plot of the standardized Pearson (chi) residual that is used to identify observations that are poorly accounted for by the model versus the predicted values.
Standardized Deviance residual	creates a plot of the Standardized deviance residual that is used to identify poorly fitted values versus the predicted values.

Generalized Linear Models: Setting Predictions Options

In the selection pane, click **Predictions** to access these options.

Option Name	Description
Data to predict	identifies the data sources that you want to use to predict the output data. If you select the **Additional data** check box, then this data is included. To designate the location of your additional data, click **Browse**.
Save output data	specifies whether to save the predictions and diagnostic statistics in an output data set: Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Additional statistics	specifies whether to save any additional statistics. ■ **Residuals** saves the likelihood residual that is used to identify poorly fitted observations, the standardized Pearson (chi) residual that is used to identify observations that are poorly accounted for by the model, and the standardized deviance residual that is used to identify poorly fitted observations. ■ **Prediction limits** saves the upper and lower confidence limits for the predicted value.
Show predictions	shows the predictions in the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

37

High-Performance Linear Regression

About the High-Performance Linear Regression Task

The High-Performance Linear Regression task fits and performs model selection for ordinary least squares models. Unlike the Linear Regression task, you can specify general linear models that include classification variables. The High-Performance Linear Regression task also supports the least angle regression and lasso selection methods.

The High-Performance Linear Regression task allows two parameterizations (GLM and reference) for the classification variables. Unlike other tasks that allow multiple parameterizations, the classification variables in this task have the same parameterization. GLM parameterization is the default.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	HPREG
Required SAS products	Base SAS, SAS High-Performance Analytics Server, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

High-Performance Linear Regression: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variables	specifies the variables to use as the dependent (response) variables. You must assign at least one variable to this role.
Explanatory effects	specifies the variables to use as the regressor (explanatory) variables. If you do not specify any regressor variables, each dependent variable is fit to a mean-only model. That is, each dependent variable is fit to a model with only a constant term.
Classification variables	specifies the classification variables to use in the analysis. They can be character or numeric variables. If they are numeric, the values must be discrete rather than continuous. These variables enter the analysis through levels to which unique values are mapped. When the values are levelized, observations that share the same value are assigned to the same level. You can specify the coding style for all the classification variables or for individual classification variables. To specify the same coding style for all the classification variables, use the options in the **Global classification variable options** area. ■ Use the **GLM** option to a less-than-full-rank reference cell coding. This parameterization is used in procedures such as GLM, MIXED, and GLIMMIX in SAS/STAT. ■ Use the **Reference** option for reference cell encoding. Specify whether you want to use the first or last ordered value as a reference. If you select **Reference**, you can override the reference cell encoding for an individual classification variable. The **Split** check box specifies which columns of the design matrix correspond to any effect that contains a split classification variable. Each column can be selected to enter or leave a model independently of the other design columns for that effect. You can specify this as a global option or as an option on an individual classification variable. Suppose that the variable Temp has three levels with values hot, warm, and cold and that the variable gender has two levels with values M and F. ■ If **Split** is selected as a global option, the two effects are split into eight independent effects. The effect gender is split into two effects: gender_M and gender_F. The effect gender*temp is split into six effects: gender_M*temp_hot, gender_F*temp_hot, gender_M*temp_warm, gender_F*temp_warm, gender_M*temp_cold, and gender_F*temp_cold ■ If **Split** is selected as an option for the temp variable, the effect gender*temp is split into three effects: gender*temp_hot, gender*temp_warm, and gender*temp_cold. Each of these three split effects now has two parameters that correspond to the two levels of gender.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.

Role Name	Description
Relative weight	specifies the variable that contains the values that are relative weights for a weighted least squares fit. The values of the relative weight variable must be nonnegative. You can assign a maximum of one variable to this role.

High-Performance Linear Regression: Building a Model

Model Options

About Models

In the selection pane, click **Model** to access these options.

By default, no effects are specified, which results in the task fitting an intercept-only model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects.

Specify Main Effects

1 Select the variable name in the **Class and quantitative variables** box.

2 Click **Main** to add the variable to the **Effects** box.

Specify Crossed Effects (Interactions)

1 Select one or more variable names. To select more than one variable, press Ctrl while you select the additional variable names.

2 Click **Cross**. If you select only one variable, then the variable is crossed with itself.

Specify Nested Effects

1 Select one or more main effects or crossed effects that you want to nest.

2 Select one or more Class variables within which you want to nest the main effects or crossed effects.

3 Click **Nest**. The nested effect appears in the **Effects** box in the form *variable 1*(*variable 2*).

Create Multiple Crossings of Two or More Selected Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Factorial** to add the factorial effects to the **Effects** box. By default, each combination of two variables is used.

Create Polynomial Crossing of Quantitative Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Polynomial** to add the polynomial effects to the **Effects** box.

Remove the Effects of the Intercept from the Analysis

Clear the **Include intercept check** box. If you clear this check box, then you must have specified at least one effect. By default, the intercept is included.

Specifying a Model Selection Method

In the selection pane under the Model heading, click **Selection** to access these options.

Option Name	Description
Model selection method	Select a method from the drop-down list of model selection methods. If you do not select a method, the full model is used. Depending on the model selection method, additional options for that method become available. Here are the available methods: ■ **Full model fitted (no selection)** is the default and provides no effect selection. This is the model that was created when you assigned the dependent variables and quantitative variables is used to fit the model. ■ **Forward selection**begins with no effects in the model. For each of the explanatory variables, this method calculates *F* statistics that reflect the variable's contribution to the model, if it is included. The *p*-values for these *F* statistics are compared to the significance level that is specified for including a variable in the model. By default, this value is `0.05`. To change this significance level, enter the value in the **To enter the model** text box. If no *F* statistic has a significance level greater than this value, the forward selection stops. Otherwise, the forward selection method adds the variable with the largest *F* statistic to the model. The forward selection method then calculates *F* statistics again for the variables that remain outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant *F* statistic. After a variable is added to the model, it stays there. ■ **Backward elimination** begins by calculating *F* statistics for a model, including all the explanatory variables. Then the variables are deleted from the model one by one until all the variables that remain in the model produce significant *F* statistics. The significance level is specified in the To stay in the model text box. By default, this value is `0.05`. At each step, the variable that shows the smallest contribution to the model is deleted. ■ **Stepwise selection** is a modification of the forward selection method. In the stepwise method, variables that are already in the model do not necessarily stay there. As in the forward selection method, variables are added one by one to the model if the *F* statistic is significant. The significance level is specified in the To enter the model text box. However, after a variable is added, the stepwise method checks all the variables that are already included in the model and deletes any variable that does not produce a significant *F* statistic. The significance level is specified in the **To stay in the model** text box. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when either of these conditions is met: □ no variable outside the model has a significant *F* statistic, and every variable in the model is significant at the significance level that is specified to stay in the model. □ the variable to be added to the model is the variable that was just deleted from it. ■ **Forward-swap selection** is an extension of the forward selection method. This method uses pairwise swaps of effects in and out of the current model that improve the selection criterion. ■ **Least angle regression** starts with no effects in the model and adds effects. The parameter estimates at any step are smaller when compared to the corresponding least squares estimates. If the model contains classification variables, then these classification variables are split. ■ **Lasso method** adds and deletes parameters based on a version of ordinary least squares where the sum of the absolute regression coefficients is constrained. If the model contains classification variables, then these classification variables are split.

Option Name	Description
Significance levels	For the forward selection and least angle regression methods, specify a significance level to enter a quantitative (explanatory) variable in the model. For the backward elimination method, specify a significance level so that a quantitative variable stays in the model. For the stepwise selection and lasso methods, specify both significance levels.
Choose the model with the best value	Uses the criterion to select (at each step of the selection process) the model that yields the best value. If the optimal value of the specified criterion occurs for models at more than one step, then the model with the smallest number of parameters is selected. If you do not specify a criterion, then the selected model is the model at the final step in the selection process. Here are the available criteria: ■ **AIC** - Akaike's information criterion ■ **AICC** - a small-sample bias that is the corrected version of Akaike's information criterion ■ **BIC** - Schwarz's Bayesian criterion ■ **ADJRSQ** - adjusted R-square criterion ■ **CP** - Mallows' C_p statistic ■ **PRESS** - predicted residual sum of squares statistic ■ **VALIDATE** - average square error over the validation data
Choose the criteria for stopping selection process	Uses the criterion to stop the selection process. If you select **None**, then the selection process stops if no suitable add or drop candidates can be found or if a size-based limit is achieved. Here are the available criteria: ■ **SL** - the significance of the score test ■ **AIC** - Akaike's information criterion ■ **AICC** - a small-sample bias that is the corrected version of Akaike's information criterion ■ **BIC** - Schwarz's Bayesian criterion ■ **ADJRSQ** - adjusted R-square criterion ■ **CP** - Mallows' C_p statistic ■ **PRESS** - predicted residual sum of squares statistic ■ **VALIDATE** - average square error over the validation data
Effects to force into the model	Select the effect that you want to add to the model. The variable is added to the **Effects to force into the model** box. You can reorder these variables by selecting the variable and clicking the arrow buttons. Note: This option is not available if you select the **Full model fitted (no selection)** , **Least angle regression**, or **Lasso method** model selection method.

High-Performance Linear Regression: Setting the Model Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Viewing details on estimates	You can select these statistics in the **Details on estimates** area: ■ **Standardized regression coefficients** produces standardized regression coefficients. A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor. ■ **Confidence limits for parameter estimates** requests the upper and lower confidence limits for the parameter estimates. By default, the 95% limits are computed.
Specifying collinearity diagnostics	You can select these diagnostics: ■ **Tolerance values for estimate** produces tolerance values for the estimates. Tolerance for a parameter is defined as $1 – R^2$, where R^2 is obtained from the regression of the parameter on all other parameters in the model. ■ **Variance inflation values** produces variance inflation factors with the parameter estimates. Variance inflation is the reciprocal of tolerance. **Note:** These options are not available if you select the Least angle regression or Lasso method model selection methods.

High-Performance Linear Regression: Setting the Partition Options

In the selection pane, click **Partition** to access these options.

Option Name	Description
Do not partition the data for model training, validation, and testing	specifies that the observations are not partitioned.
Specify proportions to use for random assignments of observations	specifies how observations in the input data set are logically partitioned into disjoint subsets for model training, validation, and testing. You can specify that proportions of the observations be randomly assigned to training and validation roles. ■ **Training percentage** displays the proportion of observations in the input data set that are randomly assigned to the training role. Observations that are not assigned to the validation or testing roles are assigned to the training role. ■ **Validation percentage** specifies the proportion of observations in the input data set that are randomly assigned to the validation role. ■ **Testing percentage** specifies the proportion of observations in the input data set that are randomly assigned to the testing role.

High-Performance Linear Regression: Setting the Performance Options

In the selection pane, click **Performance** to access these options.

Option Name	Description
Use default settings for this task	uses the default settings.
Use custom settings for this task	enables you to specify the node, grid host, installation location, and data server for the task ■ **Nodes** specifies the number of nodes in the distributed computing environment, provided that the data is not processed alongside the database. ■ **Grid Host** specifies the name of the appliance host. ■ **Installation Location** specifies the directory in which the High-Performance Analytics shared libraries are installed on the appliance. ■ **Data Server** specifies the name of the server on Teradata systems as defined through the hosts file and as used in the LIBNAME statement for Teradata.
Show detailed procedure steps	creates a table that displays a breakdown of the procedure steps.

High-Performance Linear Regression: Setting Prediction Options

In the selection pane, click **Predictions** to access these options.

Plot Name	Description
Save output data	specifies whether to permanently save predictions and diagnostic statistics in an output data set. ■ **Predictions** saves the predicted probability of each response level. For a response variable Y with three levels, 1, 2, and 3, the individual probabilities are Pr(Y=1), Pr(Y=2), and Pr(Y=3). ■ **Diagnostic statistics** saves the DFBETAS diagnostics statistic, the confidence interval displacement diagnostics (C and CBAR), the *H* matrix, and the estimates of the linear predictor with standard error estimates. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Plot Name	Description
Additional statistics	■ Select **Residuals** to save these statistics: □ the change in the deviance □ deviance residual □ Pearson (chi-square) residual □ the change in the deviance that is attributable to deleting the individual observation □ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation ■ Select **Prediction limits** to include the upper and lower limits of a confidence interval for each prediction and the expected mean of the dependent variable in the output.
Select output variables	specifies the columns to include in the output data set.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

38

High-Performance Logistic Regression

About the High-Performance Logistic Regression Task

The High-Performance Logistic Regression task fits logistic regression models for binary, binomial, and multinomial data on the SAS appliance. The High-Performance Logistic Regression task allows two parameterizations (GLM and reference) for the classification variables. Unlike other tasks that allow multiple parameterizations, the classification variables in this task have the same parameterization. GLM parameterization is the default.

The High-Performance Logistic Regression task is specifically designed to operate in the high-performance distributed environment. By default, computations are performed in multiple threads. The Logistic Regression task uses a single thread.

Note: When you open the High-Performance Logistic Regression task, you might notice a slight delay as SAS Enterprise Guide determines which variables have binary values. Once this analysis is complete, the task provides information about the variable that you assign to the **Dependent variable** role.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	HPLOGISTIC
Required SAS products	Base SAS, SAS High-Performance Analytics Server, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

High-Performance Logistic Regression: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variable	specifies the variable to use as the dependent (response) variable. You must assign only one variable to this role.
Quantitative variables	specifies the variables to use as the independent (explanatory) variables.
Classification variables	specifies the classification variables to use in the analysis. They can be character or numeric variables. If they are numeric, the values must be discrete rather than continuous.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies one variable to use as relative weights for the corresponding values of the variable that you selected as the response variable. If a variable value is negative or missing, it is excluded from the analysis.

High-Performance Logistic Regression: Building a Model

Specifying the Response

In the selection pane under the Model heading, click **Response** to access these options.

Role Name	Description
Response type	The available response types depend on the number of levels in your response variable. ■ If your response variable has only two levels, the response type is binary. ■ If your response variable has more than two levels, you can choose either ordered or unordered as the response type.

Role Name	Description
Type of model	You can select from these model types: ■ **logit** displays the output from the log odds function, which is the default selection. The binary logit model is used when there are two response categories. The cumulative logit model is used when there are more than two response categories. ■ **probit** displays the output from the inverse standard normal probability integral function. The binary probit model is used when there are two response categories. The cumulative probit model is used when there are more than two response categories. ■ **complementary log-log** displays the output from the complementary log-log function. The binary complementary log-log model is used when there are two response categories. The cumulative complementary log-log model is used when there are more than two response categories. ■ **log-log** displays the output from the log-log function. The binary log-log model is used when there are two response categories. The cumulative log-log model is used when there are more than two response categories. ■ **glogit** displays the output from the generalized logit function. In the generalized logit model, each nonreference category is contrasted with the reference category. **Note:** This option is available only if you specified **Unordered** as the response type.
Response levels	The Response levels box lists in ascending order the formatted levels for the response variable.
Fit model to level	You can specify how to fit the model to a level. ■ If you selected **Binary** as the response type, then from the Fit model to level drop-down list select the event category for the binary response model. ■ If you selected Unordered as the response type, then from the **Fit model to level** drop-down list select the reference category for the generalized logit model. Each logit contrasts a nonreference category with the reference category. ■ If you selected **Ordered** as the response type, then you can use the values in the **Fit model to level** drop-down list to specify the order of the response categories. If you select the first ordered value, then the response categories are ordered as they appear in the **Response levels** box. If you select the last ordered value, then the response categories are in reverse order.

Model Options

About Models

In the selection pane, click **Model** to access these options.

By default, no effects are specified, which results in the task fitting an intercept-only model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects.

Specify Main Effects

1 Select the variable name in the **Class and quantitative variables** box.

2 Click **Main** to add the variable to the **Effects** box.

Specify Crossed Effects (Interactions)

1 Select one or more variable names. To select more than one variable, press Ctrl while you select the additional variable names.

2 Click **Cross**. If you select only one variable, then the variable is crossed with itself.

Specify Nested Effects

1 Select one or more main effects or crossed effects that you want to nest.

2 Select one or more Class variables within which you want to nest the main effects or crossed effects.

3 Click **Nest**. The nested effect appears in the **Effects** box in the form *variable 1*(*variable 2*).

Create Multiple Crossings of Two or More Selected Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Factorial** to add the factorial effects to the **Effects** box. By default, each combination of two variables is used.

Create Polynomial Crossing of Quantitative Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Polynomial** to add the polynomial effects to the **Effects** box.

Remove the Effects of the Intercept from the Analysis

Clear the **Include intercept check** box. If you clear this check box, then you must have specified at least one effect. By default, the intercept is included.

Specifying a Model Selection Method

In the selection pane under the Model heading, click **Selection** to access these options.

Option Name	Description
Model selection method	Select a method from the drop-down list of model selection methods. If you do not select a method, the full model is used. Depending on the model selection method, additional options for that method become available. Here are the available methods: ■ **Full model fitted (no selection)** is the default and provides no effect selection. This is the model that was created when you assigned the dependent variables and quantitative variables is used to fit the model. ■ **Forward selection**begins with no effects in the model. For each of the explanatory variables, this method calculates *F* statistics that reflect the variable's contribution to the model, if it is included. The *p*-values for these *F* statistics are compared to the significance level that is specified for including a variable in the model. By default, this value is `0.05`. To change this significance level, enter the value in the **To enter the model** text box. If no *F* statistic has a significance level greater than this value, the forward selection stops. Otherwise, the forward selection method adds the variable with the largest *F* statistic to the model. The forward selection method then calculates *F* statistics again for the variables that remain outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant *F* statistic. After a variable is added to the model, it stays there. ■ **Backward elimination** begins by calculating *F* statistics for a model, including all the explanatory variables. Then the variables are deleted from the model one by one until all the variables that remain in the model produce significant *F* statistics. The significance level is specified in the To stay in the model text box. By default, this value is `0.05`. At each step, the variable that shows the smallest contribution to the model is deleted. ■ **Stepwise selection** is a modification of the forward selection method. In the stepwise method, variables that are already in the model do not necessarily stay there. As in the forward selection method, variables are added one by one to the model if the *F* statistic is significant. The significance level is specified in the To enter the model text box. However, after a variable is added, the stepwise method checks all the variables that are already included in the model and deletes any variable that does not produce a significant *F* statistic. The significance level is specified in the **To stay in the model** text box. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when either of these conditions is met: □ no variable outside the model has a significant *F* statistic, and every variable in the model is significant at the significance level that is specified to stay in the model. □ the variable to be added to the model is the variable that was just deleted from it. ■ **Backward elimination (fast with no model refitting)** performs fast backward elimination. This method starts with all effects in the model and deletes effects without refitting the model.

Option Name	Description
Significance levels	■ For the forward selection method, specify a significance level to enter a quantitative (explanatory) variable in the model. ■ For the backward elimination and backward elimination (fast with no model refitting) methods, specify a significance level so that a quantitative variable stays in the model. ■ For the stepwise selection method, specify both significance levels.
Choose the model with the best value	Uses the criterion to select (at each step of the selection process) the model that yields the best value. If the optimal value of the specified criterion occurs for models at more than one step, then the model with the smallest number of parameters is selected. If you do not specify a criterion, then the selected model is the model at the final step in the selection process. Here are the available criteria: ■ **AIC** - Akaike's information criterion ■ **AICC** - a small-sample bias that is the corrected version of Akaike's information criterion ■ **BIC** - Schwarz's Bayesian criterion
Choose the criteria for stopping selection process	Uses the criterion to stop the selection process. If you select **None**, then the selection process stops if no suitable add or drop candidates can be found or if a size-based limit is achieved. Here are the available criteria: ■ **SL** - the significance of the score test ■ **AIC** - Akaike's information criterion ■ **AICC** - a small-sample bias that is the corrected version of Akaike's information criterion ■ **BIC** - Schwarz's Bayesian criterion
Effects to force into the model	Select the effect that you want to add to the model. The variable is added to the **Effects to force into the model** box. You can reorder these variables by selecting the variable and clicking the arrow buttons. **Note:** This option is not available if you select the **Full model fitted (no selection)** model selection method.

High-Performance Logistic Regression: Setting the Model Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Model fit assessment	specifies these model fit statistics: ■ **Hosmer and Lemeshow goodness-of-fit test** displays the Hosmer and Lemeshow goodness-of-fit test for the case of a binary response model. Note: This option is available only if you assign a binary variable to the **Dependent variable** role. ■ **Generalized R-squared** displays a generalized coefficient of determination for the fitted model.
Confidence level	specifies whether to compute the confidence intervals based on individual Wald tests. You can also select a value from the **Confidence level** drop-down list to set the level of significance for the confidence intervals for the regression parameters or the odds ratios.
Suppress displaying class level information	suppresses the display of the Class Level Information table. Select **Suppress the computation of the logistic regression coefficients** to suppress the computation of the covariance matrix and the standard errors of the logistic regression coefficients. When the model contains thousands of variables, the inversion of the Hessian matrix to derive the covariance matrix and the standard errors of the regression coefficients can be time-consuming.
Classification variables coding style	specifies the parameterization method for the classification variables. ■ **GLM** is used for a less-than-full-rank reference cell coding. ■ **Reference** is used for a reference cell encoding.
Converting CLASS variables to uppercase	specifies whether to convert the values of character-valued classification variables to uppercase before levelizing them. For example, if you select this option and the values of a classification variable are 'a', 'A', and 'b', then 'a' and 'A' represent the same level, and the classification variable is treated as having only two values: 'A' and 'B'.

High-Performance Logistic Regression: Setting the Performance Options

In the selection pane, click **Performance** to access these options.

Option Name	Description
Use default settings for this task	uses the default settings.

Option Name	Description
Use custom settings for this task	enables you to specify the node, grid host, installation location, and data server for the task ■ **Nodes** specifies the number of nodes in the distributed computing environment, provided that the data is not processed alongside the database. ■ **Grid Host** specifies the name of the appliance host. ■ **Installation Location** specifies the directory in which the High-Performance Analytics shared libraries are installed on the appliance. ■ **Data Server** specifies the name of the server on Teradata systems as defined through the hosts file and as used in the LIBNAME statement for Teradata.
Show detailed procedure steps	creates a table that displays a breakdown of the procedure steps.

High-Performance Logistic Regression: Setting Prediction Options

In the selection pane, click **Predictions** to access these options.

Plot Name	Description
Save output data	specifies whether to permanently save predictions and diagnostic statistics in an output data set. ■ **Predictions** saves the predicted probability of each response level. For a response variable Y with three levels, 1, 2, and 3, the individual probabilities are Pr(Y=1), Pr(Y=2), and Pr(Y=3). ■ **Diagnostic statistics** saves the DFBETAS diagnostics statistic, the confidence interval displacement diagnostics (C and CBAR), the *H* matrix, and the estimates of the linear predictor with standard error estimates. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Plot Name	Description
Additional statistics	■ Select **Residuals** to save these statistics: □ the change in the deviance □ deviance residual □ Pearson (chi-square) residual □ the change in the deviance that is attributable to deleting the individual observation □ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation ■ Select **Prediction limits** to include the upper and lower limits of a confidence interval for each prediction and the expected mean of the dependent variable in the output.
Select output variables	specifies the columns to include in the output data set.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

39

Histograms

About the Histograms Task

A process capability analysis compares the distribution of output from a process in statistical control to its specification limits in order to determine the consistency with which the specifications are met. Histograms are used in process capability analysis to compare the distribution of measurements from a process in statistical control to its specification limits.

You might use the Histograms task to analyze a manufacturing process. For example, suppose that a semiconductor manufacturer produces printed circuit boards that are sampled to determine whether the thickness of their copper plating falls between a lower specification limit of 3.45 mils and an upper specification limit of 3.55 mils. Each bar in the histogram shows the percentage of observations and the measured thickness. The lower and upper specification limits are included in the histogram, so that you can see the percentage of observations that fall within these specification limits.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CAPABILITY

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Capability Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the numeric variables to be analyzed. You must assign at least one variable to this role. For all plots except the P-P plot, you can specify an upper limit, a target value, and a lower limit in the **Spec limits for *variable-name*** area for each analysis variable. When you specify these limits, the output automatically includes a table that shows the tests for normality, a table that shows the specification limits, and a table that shows the process capability indices.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. Note: To prevent the input data set from being sorted, clear the **Sort by variables** check box. Note: You cannot group analyses by a variable that you have already selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight. When you assign a variable to this role, the value of the variable in each observation is used to compute weighted statistics. You can assign only one variable to this role. Note: Assigning a variable to the **Relative weight** role is not recommended when you are creating a histogram, probability plot, P-P plot, Q-Q plot, or CDF plot. If you assign a variable to this role, any output should be interpreted with caution.
Classification variables (available for the Histogram task only)	specifies one or two variables to use as classification variables. When you assign a variable to this role, the task creates a comparative histogram that enables you to compare the distribution of an analysis variable across classification levels. Note: To create comparative probability plots or comparative Q-Q plots, use the Distribution Analysis task.

Histograms: Selecting the Distribution

In the selection pane under the Distributions heading, click **Summary** to access these options.

You can fit several theoretical distributions on a single histogram. After selecting the distribution to compute, select the distribution name in the selection pane to specify the parameter values and line properties for that distribution.

Option Name	Description
Distributions	Use the check boxes to select one or more of these estimated density curves: ■ **Normal** fits a normal density with a mean (mu) and a standard deviation (sigma). ■ **Lognormal** fits a lognormal density with a scale parameter (zeta), a threshold parameter (theta), and a shape parameter (sigma). ■ **Exponential** fits an exponential density with a scale parameter (sigma) and a threshold parameter (theta). ■ **Weibull** fits a three-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). ■ **Beta** fits a beta density with a scale parameter (sigma), a threshold parameter (theta), and shape parameters (alpha and beta). ■ **Gamma** fits a gamma density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (alpha). ■ **Kernel** fits nonparametric kernel density estimates. Note: For comparative histograms, only the **Normal** distribution and the **Kernel** density estimates are available.
Graphics style	Select the graphics style to use in the output. You can choose between traditional SAS graphics or the ODS statistical graphics. For more information about ODS graphics, see *SAS Output Delivery System: User's Guide*.

Histograms: Customizing the Distribution

How to Access the Distribution Options

In the selection pane under the Distributions heading, click the name of the distribution that you want to customize to access these options.

Note: These distribution options are available only if you previously selected the distribution on the **Summary** page or if you select the ***Distribution name*** check box on the page for that distribution.

Specifying the Parameters of the Distribution

For the normal, lognormal, exponential, Weibull, beta, and gamma distributions, you can specify the parameters for each analysis variable or all of the analysis variables. By default, the task estimates the parameter values for all of the analysis variables.

To specify the parameters for each analysis variable:

1. Clear the **Apply distribution on all variables** check box.
2. Select the variable's check box in the **Analysis variables** box.
3. Specify the parameter values.

Specifying Kernel Density Estimates

For nonparametric kernel density estimates, you can specify the bandwidth (c) and the kernel function (k). By default, the task estimates a bandwidth that minimizes the approximate mean integrated square error (MISE). If you do not want to estimate this bandwidth, then clear the **Minimize MISE** check box. To specify the standardized bandwidth parameter for kernel density, enter the bandwidth in the text box. You can specify up to five values with a space between each value. For example, enter these values in the text box: `0.5 1.0 1.5`.

Note: If you do not clear the **Minimize MISE** check box, then MISE counts as one of the five bandwidth values that you can specify.

You can also specify the kernel function (normal, quadratic, or triangular) that is used to compute the kernel density estimate. The default kernel function is **Normal**.

You can specify kernel functions for up to five estimates. If you specify more than one kernel function, the order of the functions is always normal, quadratic, triangular. To add functions or change the order, you must edit the task code.

If you specify more kernel functions than bandwidths, then the last bandwidth is repeated for the remaining estimates. For example, if you specify 0.5 and 1.0 as the bandwidths and specify normal, quadratic, and triangular as the kernel functions, then the 1.0 bandwidth is repeated for the triangular function. If you specify more bandwidths than kernel functions, then the last kernel function is repeated for the remaining bandwidths.

For example, if you specify 0.5, 1.0, and 1.5 as the bandwidths and specify normal and quadratic as the kernel functions, the quadratic function is used for the 1.5 bandwidth.

Capability Analysis: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the color of the tick marks

- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Adding an Inset

In the selection pane under the **Appearance** heading, click **Inset** to access these options.

You can enhance a plot by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative histogram, an inset appears in each component histogram.

Select **Include inset** to add summary statistics to the plot. Then select the statistics that you want to include in the inset.

Note: The statistics for calculating the capability indices are available only if you created specification limits for the variable that you assigned to the **Analysis variables** role. You can create specification limits for histograms, CDF plots, probability plots, and Q-Q plots. For more information about calculating capability indices, see the Help for the CAPABILITY procedure.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

The plot appearance options that are available depend on which type of plot the task produces.

Role Name	Description
Background	specifies the background color for the plot.
Bar outline and **Bar fill**	specifies the colors for the bars in a histogram. **Note:** These options are available only for histograms.
Reverse the two axes putting the percentiles on the opposite axis	reverses the axes on the plot. If you select this option, then the axes options and reference lines that you selected might no longer work since the orientation of the plot will have changed. **Note:** This option is available only for probability plots and Q-Q plots.
Suppress default legends	prevents the default legend from being displayed with the plot.

Role Name	Description
Use midpoints	specify midpoints for the histogram. The options that are available from the drop-down list depend on whether you are creating a histogram or a comparative histogram. If you assign a variable to the **Classification variables** role, you are creating a comparative histogram. For comparative histograms, these options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Key** determines the midpoints for the data in the key cell. By default, the levels for the classification variable are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first classification variable. If you specified only one classification variable, then the key cell is the level that occurs first in this order. If you specified two classification variables, then the key cell is the combination of levels of variable 1 and variable 2 that occurs first in this order. Thus, the choice of the key cell determines the uniform horizontal axis that is used for all cells. The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. ■ **Uniform** determines the number of midpoints based on the total sample size. If you do not assign a variable to the **Classification variables** role, you are creating a histogram. For histograms, the following options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Specify number of bins** enables you to specify the number of bins (also called histogram intervals) for the data. By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. **Note:** You can specify midpoints only if you are using the Histograms task.
Specify custom chart size (in pixels)	specifies the image size. For reference, the total screen size is given in the **Width** and **Height** boxes.

Note: Comparative options are available for comparative histograms. You can create a comparative histogram by assigning a variable to the **Classification variables** role in the Histograms task.

Role Name	Description
Rows and **Columns**	If you assign at least one classification variable to an analysis role, then you can use the **Rows per page** and **Columns per page** boxes to change the arrangement of rows and columns in the comparative histogram. If you assign one classification variable, the default arrangement is two rows and one column per page. If you assign two classification variables, the default arrangement is two rows and two columns per page. You can also use the **Frame side** and **Frame top** drop-down lists to specify the colors that are used to fill the frame area of the row labels and the column labels.
Number of bins	By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. You can also specify a standard deviation such that the number of bins that are displayed is based on a range of the standard deviation value above and below the mean of the data.

Setting Specification Limit Options

Note: These options are available for histograms, probability plots, Q-Q plots, and CDF plots.

In the selection pane under the **Appearance** heading, click **Spec Limits** to specify the lower specification limit, the target, and the upper specification limit.

Note: The **Spec Limits** options are available only if you specify one or more limits for at least one analysis variable. To specify a specification limit for an analysis variable, click **Data** in the selection pane, and then select the analysis variable.

Capability Analysis: Specifying the Types of Analyses

In the selection pane, click **Tables** to access these options.

By default, the Capability Analysis report includes the following tables: a basic confidence intervals table, a basic measures table, a tests for location table, and a moments table.

To add a table to the report , select the table in the **Tables** box.

To specify the analysis options, select the table in the **Tables** box. If options are available, a box appears on the right that you use to specify the options.

To suppress all descriptive statistics tables from the report , select the **Suppress descriptive statistics and capability indices tables** check box. Selecting this option does not suppress these tables:

- tables that are created by the distribution that you selected. You can suppress these tables by selecting the **Suppress distribution tables** check box on the panel for that distribution.
- tables that are created by the INTERVALS statement. To suppress these tables, you can specify the NOPRINT option. For more information about the INTERVALS statement, see the Help for the CAPABILITY procedure.

To save your results to an output data set , select the **Save output statistics to a data set** check box. The data set will include univariate statistics and capability indices.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

40

Individual Measurements Chart

About the Individual Measurement Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The Individual Measurements Chart task creates control charts for the individual measurements and the moving ranges. These charts are appropriate when only one measurement is available for each subgroup sample and when the measurements are independently and normally distributed.

You might want to use this task to analyze a manufacturing process. For example, suppose that an aeronautics company that manufactures jet engines measures the inner diameter of the forward face of each engine (in centimeters). The diameter measurements of 20 engines are stored in a SAS data set. Each point on the individual measurements chart indicates the inner diameter of a particular engine. Each point on the moving range chart indicates the range of the two most recent measurements. If all the individual measurements and moving ranges fall within the control limits, you can conclude that the process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Individual Measurements Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Process measurement	specifies the process variables to be analyzed. These variables contain the individual measurements. You must assign at least one variable to this role.
Measurement identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A measurement identifier is required, and you assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.

Option Name	Description
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.

Role Name	Description
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

41

Life Tables

About the Life Tables Task

The Life Tables task computes nonparametric estimates of the survival distribution of data that might be right-censored due either to withdrawals or to termination of the study. For such data, you know only that the lifetime exceeded the supplied value. Usually, you want to compare survival curves to determine whether two samples come from the same survival distribution.

This task computes rank tests and a likelihood ratio test to test for homogeneity of survival functions across strata. The task also generates plots that enable you to examine the survival curves.

You might want to use this task to study the survival distribution for a specific population. For example, suppose that you want to study the survival time of 40 rats that were exposed to a carcinogen in a small, randomized trial and assigned to one of two treatment groups. The survival time is the time from randomization to death. The event of interest is death from cancer that was induced by the carcinogen. You can use the Life Tables task to determine whether the survival distributions differ between the two treatment groups.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	LIFETEST
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Life Tables: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Survival time	specifies the variable to use as the survival time variable. Survival times are often called failure times, and event times are uncensored survival times. Survival time is required, and you can assign only one variable to this role.
Censoring variable	specifies the variable to use as the censoring variable. In the **Right Censoring Values** area, select the values of this variable that indicate whether a survival time is censored. The values of the censoring variable should be numeric, nonmissing values. The censoring variable is not required. However, if it is used, you can specify only one variable, and you must indicate the censoring values.
Strata variables	specifies the variables whose values determine the stratum levels. If you do not specify endpoints for the numeric variables, then the strata are determined by the combination of levels of the unique values of the strata variables. You can click on the check box to allow missing values to form valid stratum levels. For numeric strata variables, you can specify the strata intervals by selecting **Specify intervals** in the **Stratum levels** area. Then define the intervals by adding endpoints. By default, the unique values form the strata levels. Specify the endpoints that determine stratification. The interval below an endpoint does not contain the endpoint, and the interval above the endpoint does contain the endpoint. To specify endpoints for a variable, enter endpoint values and click **Add**. The endpoints are not limited to integer values. You can enter single values separated by commas or spaces, or you can enter ranges. For example, enter `5 10 15` to produce unique strata for values that fall into the intervals { `[-infinity,5)`, `[5,10)`, `[10,15)`, `[15,infinity)` }. The range `5 ... 10` produces unique strata for values that fall into the following intervals: { `[-infinity,5)`, `[5,6)`, `[6,7)`, `[7,8)`, `[8,9)`, `[9,10)`, `[10,infinity)` }. If you assign multiple variables to this role, define the intervals separately for each variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Test covariates	specifies a list of numeric (continuous) covariates that are used to test for association with the failure time. Two sets of rank statistics are computed. The statistics and their variances are pooled over all strata. Univariate (marginal) test statistics are displayed for each of the covariates. Test statistics are also displayed for joint effects of covariances.

Role Name	Description
ID variables	specifies one or more variables to use as ID variables. The values of ID variables are used to label the observations of the product-limit survival estimates.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you already selected for an analysis role.

Life Tables: Setting Method Options

In the selection pane, click **Method** to access these options.

Option Name	Description
Estimation method	To specify an estimation method, click either the **Product-limit (Kaplan-Meier)** or **Life table** button. Both methods generate similar results. However, the product-limit method (also called the Kaplan-Meier method) is based on actual survival times, and the life table method groups survival times into intervals. When the sample size is very large, it might be useful to group the survival times into intervals and perform the life table analysis. The default method is product-limit.
Intervals	If you select **Product-limit** as the estimation method, then you can reduce the number of survival estimates that are displayed by defining endpoints for each interval. The task displays only the estimates for the smallest time within each specified interval. If you select **Life table** as the estimation method, then you can specify information about the construction of the intervals that are used in the calculation of the survival function. By default, 10 intervals are used. You can specify a different value by selecting **Number** and then entering your value. You can also construct intervals by specifying the width for the interval. Select **Width** and enter your value. If you want to define exact endpoints to use for constructing the intervals, select **Endpoints**.
Confidence level for survivor function	You can either enter a value for the confidence level, or you can select one from the drop-down list. The default value is a 95% confidence level.

Life Tables: Generating Plots

In the selection pane, click **Plots** to access these options.

Use these options to produce plots of estimated survival functions, hazard functions, and probability density functions versus time.

Option Name	Description
Show survival function plot	This option generates a plot of the survival distribution function, which is used to describe the lifetime of the population of interest. ■ To include confidence limits, select the **Include pointwise confidence intervals** check box. The confidence limits for the plots are controlled by the Methods setting for the confidence intervals. The default value is a 95% confidence interval. ■ To show all of the plots for different strata on a single plot with a legend, select the **Overlay strata on single plot** check box. If you assigned a variable to the **Strata variables** role and the **Show survival function plot** check box is selected, then all of the plots for different strata are on a single plot by default. Note: The **Overlay strata on single plot** check box is activated only when one variable is assigned to the **Strata variables** role. If more than one variable is assigned, then you cannot select this option.
Probit(1 — survival function)	This option generates a plot of the probit of 1 minus the estimated survival function versus the log of time. This type of plot is useful for evaluating lognormal distributions. If the lognormal model is appropriate, then the curve should be approximately linear.
-Log(survival function)	This option generates a plot of the negative log of the estimated survivor function versus time. This type of plot is useful for evaluating exponential distributions. If the exponential model is appropriate, then the curve should be approximately linear through the origin.
Log(-log(survival function))	This option generates a plot of the log of the negative log of the estimated survivor function against the log of time. This type of plot is useful for evaluating a Weibull distribution. If the Weibull model is appropriate, then the curve should be approximately linear.
Life-table estimation	To generate plots of the hazard function and the probability density function versus time, you must first select the life table method to compute the survival function estimates. Then you can select the **Hazard function** or **Probability density function** check box.

Life Tables: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Save output data	creates output data sets that contain the survival estimates plus various statistics. ■ Select the **Survival estimates** check box to create an output data set that contains the estimates of the survival function and the corresponding confidence limits for all strata. ■ Select the **Statistics** check box to create an output data set that contains the overall chi-square test statistic for association with failure time for the variables in the **Test covariates** roles that you selected. This data set also includes the values of the univariate rank statistics and their estimated covariance matrix. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress all displayed output	Does not display the results. If you select this check box, then the options on the Plots panel are not available. The results are displayed by default.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified

- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

42

Line Plot

About the Line Plot Task

The Line Plot task creates line, scatter, spline, needle, step, regression, smooth, STD, Lagrange interpolation, or overlay plots that show the mathematical relationships between variables by revealing trends or patterns of data points.

For example, a line plot can show whether there is a relationship between advertising expenditures and sales for a specific product.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GPLOT
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Line Plot: Selecting a Plot Type

In the selection pane, click **Line Plot** to access these options.

Role Name	Description
Line Plot	creates a plot where the data points are connected with straight lines. Points are connected in the order in which they occur in the input data set.
Spline Plot	specifies that a spline routine be used for interpolating the line.
Needle Plot	draws a vertical line from each data point to a horizontal line at the 0 value on the vertical axis or the minimum value on the vertical axis if it is greater than 0. The horizontal line is drawn automatically.
Step Plot	specifies that the data be plotted with a step function. In the plot, the steps are joined with a vertical line. Before plotting, the data is sorted by the independent variable.
Scatter Plot with Regression Line	creates a scatter plot and fits the best linear regression line for the data. By default, regression lines are not forced through plot origins and confidence limits are not displayed.
Smooth Plot	specifies that a smooth line be fit to data by using a spline routine. This interpolation method is used for smoothing noisy data. The points on the plot do not necessarily fall on the line. Before plotting, the data is sorted by the independent variable.
Standard Deviation Plot	specifies that a solid line connect the mean value of the Y variable with two standard deviations for each value of the X variable.
Lagrange Interpolation Plot	specifies that a Lagrange polynomial of the first degree be used to smooth the plot line. Before plotting, the data is sorted by the independent variable.
Multiple vertical column line plots using overlay	creates a separate plot for each Y variable versus the same X variable. All of the plots appear on the same graph.
Multiple line plots by group column	creates a separate plot for each distinct value of the group variable. All of the plots appear on the same graph.
Scatter Plot	creates a scatter plot that shows the relationships between two or three variables by revealing patterns or concentrations of data points.

Line Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart. You can create a statistical summary of the unique values of the horizontal variable. Select the variable under the **Vertical** role. Select **Summarize for each distinct horizontal value** and select the summary function. For example, your input data lists the sales and expenses for each country by region. You assign the Country variable to the **Horizontal** role, and the Sales variable to the **Vertical** role. If you want to plot the sum of the sales for each country, select the check box, and then select **Sum** from the drop-down list. The task adds the regional sales for a country and presents a single sales value for each country in the final output.
Vertical (Right)	The column that you assign to this role is the vertical or Y axis variable for the right side of the graph. This variable is plotted against the X axis variable so that another plot is produced on the graph. You can create a statistical summary of the unique values of the horizontal variable. Select the variable under the **Vertical** role. Select **Summarize for each distinct horizontal value** and select the summary function. For example, your input data lists the sales and expenses for each country by region. You assign the Country variable to the **Horizontal** role, and the Expenses variable to the **Vertical (Right)** role. If you want to plot the sum of the expenses for each country, select the check box, and then select **Sum** from the drop-down list. The task adds the regional expenses for a country and presents a single expense value for each country in the final output. **Note:** The **Vertical (Right)** role is not available if you select **Multiple line plots by group column** as the type of line plot.
Group	The values of the column that you assign to this role determine how many plot lines are drawn on the same graph. A separate plot is created for each unique value of this column. All of the lines use the same set of axes. **Note:** The **Vertical (Right)** role is not available if you select **Multiple line plots by group column** as the type of line plot.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Select the **Produce same axis scaling for all graphs** check box if you want the scale for each axis to be the same for all graphs that are produced. By default, the range of values for each axis is based on the minimum and maximum values in the data that is being plotted, and therefore it could vary from graph to graph. If you select this option, then the range for each axis is the same for all the graphs.

Line Plot: Setting Appearance Options

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Plots** to access these options.

Note: These options are available only if you select a line plot or two-dimensional scatter plot.

To control the appearance of the plot for a specific variable:

1 Select a variable from the list. The variables in the list are the variables that you assigned to the **Vertical** and **Vertical (Right)** roles. If you selected **Multiple line plots by group column** as the type of line plot, this list contains the unique values of the variable that you assigned to the **Group** role.

2 In the **Line** area, specify the line style, the line width, and the color of the line.

3 Select an outline color. This option specifies the color for these parts:

- the outline of any enclosed areas
- confidence limits in a regression analysis
- the lines or outlines for standard deviation plots

4 In the **Data point marker** area, specify the type of marker, the symbol to use, the height of each marker, and the color for the data point markers.

Note: Some symbols are not supported by all graphics output formats. If you select a symbol that is not supported by the output format, then a different symbol (one that is supported) is displayed instead.

5 Select whether to break the plot lines at missing values.

Note: The **Break the second plot line at missing values** check box is available only if you assign a variable to the **Vertical (Right)** role.

Selecting the Interpolation Method

In the selection pane under the Appearance heading, click **Interpolations** to access these options.

To select an interpolation method:

1 Select a variable. If you are creating a line plot or a two-dimensional scatter plot, the variables in the list are the variables that you assigned to the **Vertical** and **Vertical (Right)** roles. If you selected **Multiple line plots by group column** as the type of line plot, the list contains the unique values of the variable that you assigned to the **Group** role.

2 Select an interpolation method from the drop-down list. Depending on the interpolation method that you choose, you might need to specify additional settings. These interpolation methods are available:

- **LaGrange** draws a smooth curve through the data points. This method is used primarily when data consists of tabulated, precise values. A polynomial of the specified degree (1, 3, or 5) is fitted through the nearest 2, 4, or 6 points. In general, the first derivative is not continuous. If the values of the X variable are not strictly increasing, the corresponding parametric method is used.
- **Line** connects data points with straight lines. Points are connected in the order in which they occur in the input data.
- **Needle** draws a vertical line from each data point to a horizontal line at the 0 value on the vertical axis or the minimum value on the vertical axis if it is greater than 0. The horizontal line is drawn automatically.
- **Regression** creates a regression analysis plot. This interpolation method is available only if you assign numeric variables to the Horizontal and Vertical roles.
- **Scatter** suppresses any interpolation. A plot of the data points is created.
- **Smooth** fits a smooth line to the data by using a spline routine. This is a method for smoothing noisy data. The points on the plot do not necessarily fall on the line.
- **Spline** plots the line by using a spline routine.
- **STD** uses a solid line to connect the mean Y value with ± 1, 2, or 3 standard deviations for each X. By default, two standard deviations are used.
- **Step** plots the data by using a step function.

3 (Optional) Select the **Apply to all** check box to use the same interpolation method and property settings for all the variables in the list. This option is available only if there is more than one variable in the list.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. **Note:** This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. **Note:** If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.

Role Name	Description
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

43

Line Plot Wizard

About the Line Plot Wizard

The Line Plot wizard helps you create two-dimensional line plots, with options for markers and line interpolations.

Line Plot Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to assign variables to roles.

Line Plot Wizard: Assign Variables to Roles

1 Before you can assign variables to roles, you must select the data.

2 Assign variables to the following roles:

- **Horizontal** - The column that you assign to this role is the horizontal or X-axis variable for the plot.
- **Vertical** – The column that you assign to this role is the vertical or Y-axis variable for the plot.

3 (Optional) Assign a variable to the **Chart by** role. The values of the column that you assign to this role determine how many line plots to produce. A separate plot is created for each unique value. You can assign multiple variables to this role.

4 (Optional) Select the **Summarize for each distinct horizontal value** check box to create a statistical summary of the unique values of the horizontal variable. Select the summary function from the **Function** drop-down list.

For example, your input data lists the sales and expenses for each country by region. You assign the Country variable to the **Horizontal** role, and the Sales variable to the **Vertical** role. If you want to plot the sum of the sales for each country, select the check box, and then select **Sum** from the drop-down list. The task adds the regional sales for a country and presents a single sales value for each country in the final output.

5 Click **Next** to specify the plot options.

Line Plot Wizard: Set the Plot Options

1 Before you can specify the plot options, you must assign variables to roles.

2 Select the line interpolation from the **Type** drop-down list. Click **Detail** to specify any options for the interpolation that you selected. These interpolations are available:

- **LaGrange** draws a smooth curve through the data points. This method is used primarily when data consists of tabulated, precise values. A polynomial of the specified degree (1, 3, or 5) is fitted through the nearest 2, 4, or 6 points. In general, the first derivative is not continuous. If the values of the X variable are not strictly increasing, the corresponding parametric method is used.
- **Line** connects data points with straight lines. Points are connected in the order in which they occur in the input data.
- **Needle** draws a vertical line from each data point to a horizontal line at the 0 value on the vertical axis or the minimum value on the vertical axis if it is greater than 0. The horizontal line is drawn automatically.
- **Regression** creates a regression analysis plot. This interpolation method is available only if you assign numeric variables to the Horizontal and Vertical roles.
- **Scatter** suppresses any interpolation. A plot of the data points is created.
- **Smooth** fits a smooth line to the data by using a spline routine. This is a method for smoothing noisy data. The points on the plot do not necessarily fall on the line.
- **Spline** plots the line by using a spline routine.
- **STD** uses a solid line to connect the mean Y value with ± 1, 2, or 3 standard deviations for each X. By default, two standard deviations are used.
- **Step** plots the data by using a step function.

3 Specify the line style, the line width, and the color of the line.

4 Specify the symbol to use, the height of each marker, and the color for the data point markers.

Note: Some symbols are not supported by all graphics output formats. If you select a symbol that is not supported by the output format, then a different symbol (one that is supported) is displayed instead.

5 Click **Next** to specify the appearance options.

Line Plot Wizard: Specify the Appearance Options

1 Before you can set the plot appearance, you must specify the plot options.

2 Specify whether to include reference lines and tick marks on the plot.

3 (Optional) Edit the labels for the vertical and horizontal axes.

4 Click **Next** to specify the titles and footnotes.

Line Plot Wizard: Specify the Titles and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1 Before you can specify the titles and footnotes, you must set the appearance options.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

3 Click **Finish** to run the task.

44

Linear Models

About the Linear Models Task

The Linear Models task uses the method of least squares to fit general linear models. The task analyzes data within the framework of general linear models. The task handles models that relate one or several continuous dependent variables to one or several independent variables. The independent variables can be either classification variables, which divide the observations into discrete groups, or continuous variables.

You might want to use this task to perform these tasks:

- perform a regression analysis. For example, a car could be tested for gas mileage at various speeds to determine at what speed the car achieves the highest gas mileage. A quadratic model is fit to the experimental data.
- analyze a multivariate repeated measures design. For example, suppose that two responses, Y1 and Y2, are each measured three times for each subject (before treatment, after treatment, and in a later follow-up). Each subject receives one of three treatments: A, B, or the control. In this task, you can identify the different responses and another repeated factor to identify the different measurement times. The repeated measures

analysis includes multivariate tests for time and treatment main effects, as well as their interactions, across responses.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GLM
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Linear Models: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variables	specifies the variables to use as the dependent (response) variables. You must assign at least one variable to this role.
Quantitative variable	specifies the variables to use as the continuous independent (explanatory) variables (also known as effects).
Classification variables	specifies the variables to use as the discrete independent effects. Variables that you assign to this role can be numeric or character, but they should have a limited number of discrete values.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies one variable to use as relative weights for the corresponding values of the variable that you selected as the response variable. You can assign a maximum of one variable to this role.

Linear Models: Building a Model

In the selection pane, click **Model** to access these options.

By default, no effects are specified, which results in the task fitting an intercept-only model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects.

Specify Main Effects

1. Select the variable name in the **Class and quantitative variables** box.
2. Click **Main** to add the variable to the **Effects** box.

Specify Crossed Effects (Interactions)

1. Select one or more variable names. To select more than one variable, press Ctrl while you select the additional variable names.
2. Click **Cross**. If you select only one variable, then the variable is crossed with itself.

Specify Nested Effects

1. Select one or more main effects or crossed effects that you want to nest.
2. Select one or more Class variables within which you want to nest the main effects or crossed effects.
3. Click **Nest**. The nested effect appears in the **Effects** box in the form *variable 1*(*variable 2*).

Create Multiple Crossings of Two or More Selected Variables

1. Specify higher-degree crossings by adjusting the number in the **Degrees** field.
2. Click **Factorial** to add the factorial effects to the **Effects** box. By default, each combination of two variables is used.

Create Polynomial Crossing of Quantitative Variables

1. Specify higher-degree crossings by adjusting the number in the **Degrees** field.
2. Click **Polynomial** to add the polynomial effects to the **Effects** box.

Remove the Effects of the Intercept from the Analysis

Clear the **Include intercept check** box. If you clear this check box, then you must have specified at least one effect. By default, the intercept is included.

Linear Models: Setting Model Options

In the selection pane, click **Model Options** to access these options.

Option Name	Description
Show the tests associated with intercept	shows the results from hypothesis tests that were performed on the intercept.
Sum of squares to show	displays the sums of squares that are associated with the estimable functions for each effect. Estimable functions are linear combinations of the parameters of the model. Sums of squares are computed for hypothesis testing. Type I and Type III sums of squares are the defaults. You can add or remove a type of sum of squares by selecting the corresponding check box. **Note:** At least one **Sum of squares to show** check box should be selected.
Show parameter estimates	displays the parameter estimates. A parameter is a population quantity that is usually fixed but unknown, so it is predicted from the sample data by using a model. The parameter estimates are displayed by default. You can clear the **Show parameter estimates** check box only when a variable is assigned to the **Classification variables** role. Select the **Confidence limits for parameter estimates** check box to display the confidence intervals for the parameter estimates. A confidence interval is an upper and lower bound that provides a range of reasonable values for the parameter that is being estimated with a prespecified level of confidence. You can either enter a value for the confidence level or select one from the drop-down list. The default value is a 95% confidence level.

Linear Models: Setting Advanced Options

In the selection pane, click **Advanced Options** to access these options.

Option Name	Description
Estimable functions for each model effect	■ Select the **Show general form for all estimable functions** check box to show the matrix coefficients of the Type I - Type IV estimable functions for each model effect. ■ Select the check box next to the sum of squares type to show the matrix coefficients for a specific estimable function. ■ Select the **Show estimable functions as aliasing structure** check box to display the estimable functions as an aliasing structure. In an aliasing structure, each row shows which linear combination of parameters is estimated by each estimable function, and a column of the same information is added to the table of parameter estimates. **Note:** You can also show the estimable functions for the intercept if you select the **Show the tests associated with the intercept** check box.

Option Name	Description
Matrices to show	■ Select the **X'X** check box to show the augmented X'X crossproducts matrix. ■ Select the **X'X inverse** check box to show the augmented inverse (or generalized inverse) X'X matrix.
Tuning parameters	**Note:** These tuning parameters are useful only when problems are ill-conditioned. ■ To change the sensitivity in forming Type III or IV functions, enter the new value in the **Sensitivity for Type III or IV functions** box. Any element in the estimable function basis with an absolute value less than the specified value is set to 0. The default is `1E-08`. ■ To change the sensitivity that is used when sweeping a matrix, enter the desired value in the **Singularity check value** box. This option tunes the sensitivity of the regression routine to linear dependencies in the design. If a diagonal pivot element is less than the corrected sum of squares for the variable multiplied by the value that you specified, then the associated column is declared to be linearly dependent upon previous columns, and the associated parameter is set to 0. The default value is `1E-07`. ■ To display the tolerances that are used when tuning the sensitivity of the regression routine based on the value that is specified in the **Singularity check value** box, select the **Show tolerances** check box.

Linear Models: Selecting Post Hoc Tests

About the Post Hoc Tests

In the selection pane under the Post Hoc Tests heading, click **Least Squares** or **Arithmetic** to access these options.

1. Click **Add** to add an effect to the **Effects to estimate** box.
2. For each effect in the list, specify options in the **Options for means tests** box.

   Note: The options in the **Options for means test** box differ for least squares means and arithmetic means.

Specifying the Means Tests Options for Least Squares Means

The following options are available in the **Options for means tests** box.

Option Name	Description
Class effects to use	Select the class effects to use in the analysis. By default, no class effects are used.

Option Name	Description
Comparisons	Specify the comparison options to use in the analysis. ■ **Show p-values for differences** calculates the *p*-values for the differences of the least squares means. The following options are available: □ **None** does not calculate the *p*-values. □ **Default** calculates the pairwise differences with no adjustment. □ **All pairwise differences** calculates all pairwise differences. □ **Control using first level** calculates the differences with a control that, by default, is the first level of each of the specified least squares mean effects. □ **Test for less than control** tests whether the noncontrol levels are less than the control; you declare a noncontrol level to be significantly less than the control if the associated upper confidence limit for the noncontrol level minus the control is less than zero, and you ignore the associated lower confidence limits (which are set to minus infinity). □ **Test for greater than control** tests whether the noncontrol levels are greater than the control; you declare a noncontrol level to be significantly greater than the control if the associated lower confidence limit for the noncontrol level minus the control is greater than zero, and you ignore the associated upper confidence limits (which are set to infinity). ■ **Adjustment method for comparison** requests a multiple comparison adjustment for the *p*-values and confidence limits for the differences of least squares means.
Confidence limits	Select **True** for the **Show confidence limits** option to display the confidence limits for each least squares means. If you specify a value for the **Show p-values for differences** option, the confidence limits for differences between means are also shown.

Specifying the Means Tests Options for Arithmetic Means

These options are available in the **Options for means tests** box.

Option Name	Description
Class effects to use	Select the class effects to use in the analysis. By default, no class effects are used.

Option Name	Description
Comparisons	Specify the comparison method to use in the analysis: ■ **Default** specifies that no comparison method is performed. ■ **Bonferroni t-tests** performs Bonferroni *t* tests of differences between means for all main effect means. ■ **Duncan multiple range** performs Duncan's multiple range test on all main effect means ■ **Gabriel multiple com** performs Gabriel's multiple-comparison procedure on all main effect means ■ **Pairwise t-tests** performs pairwise *t* tests, equivalent to Fisher's least-significant-difference test in the case of equal cell sizes, for all main effect means ■ **Ryan-Einot-Gabriel** performs the Ryan-Einot-Gabriel-Welsch multiple range test on all main effect means ■ **Scheffe's multiple** performs Scheffé's multiple-comparison procedure on all main effect means ■ **Sidak's adjusted pai** performs pairwise *t* tests on differences between means with levels adjusted according to Sidak's inequality for all main effect means ■ **Student-Newman-Keuls** performs the Student-Newman-Keuls multiple range test on all main effect means ■ **Tukey's studentized** performs Tukey's studentized range test on all main effect means ■ **Walter-Duncan K-ratio** performs the Waller-Duncan *k*-ratio *t* test on all main effect means.
Error mean square	Select the error effect to use in the multiple comparisons. You can also specify the type of mean square for the error effect.
Means options	Specify these options for the means: ■ **Show means for** specifies whether to display the means for all of the model variables or only the dependent variables. By default, the task displays the means for all of the continuous variables. ■ **Join nonsignificant subsets** lists the means that were calculated from the comparison method in descending order and indicates nonsignificant subsets by line segments beside the corresponding means. ■ **Sort means in descending order** sorts the means in descending order. If you select **True** for the Sort means in descending order option and you select **True** for either of the options in the **Confidence intervals** area, the means are not sorted.
Confidence intervals	Specify whether to include confidence intervals in your analysis. These options are available: ■ **Show for the means** presents results from the comparison method as intervals for the mean of each level of the variables specified. ■ **Show for all pairwise differences** presents results of the comparison method as confidence intervals for all pairwise differences between means.

Option Name	Description
Homogeneity of variance	Specify whether to request a homogeneity of variance test for the groups defined by the means effect. These options are available: ■ **None** specifies that no test is performed. This is the default. ■ **Bartlett** specifies Bartlett's test, a modification of the normal-theory likelihood ratio test. ■ **Brown and Forsythe** specifies Brown and Forsythe's variation of Levene's test. ■ **Levene (square)** uses squared residuals in Levene's test. ■ **Levene (absolute)** uses absolute residuals in Levene's test. ■ **O'Brien** specifies O'Brien's test, which is basically a modification of Levene's test using squared residuals.

Linear Models: Generating Plots

In the selection pane, click **Plots** to access these options.

By default, all appropriate plots for the current data selection are included in the output. However, you can choose which plots to include in the output by selecting the **Custom lists of plots** option. You can choose from the following options:

Option Name	Description
Diagnostic plots	creates a panel of regression diagnostics for the fit be displayed. The panel displays scatter plots of residuals, absolute residuals, studentized residuals, and observed responses by predicted values; studentized residuals by leverage; Cook's *D* by observation; a Q-Q plot of residuals; a residual histogram; and a residual-fit spread plot.
Residual plots	creates scatter plots of the residuals against each continuous covariate be displayed.
Box plot	creates a plot for the model effect in a one-way analysis of variance model. To create this plot, your model can contain only one classification variable. If you did not assign a classification variable, then the box plot is not created. If you assigned a relative weight variable, these weights are ignored when creating the box plot.
Interaction plot	creates a plot for a two-way analysis of variance model. To create this plot, your model can contain only two classification variables. If you did not assign two classification variables, then the interaction plot is not created.

Linear Models: Setting Predictions Options

In the selection pane, click **Predictions** to access these options.

Option Name	Description
Data to predict	identifies the data sources that you want to use to predict the output data. If you select the **Additional data** check box, then this data is included. To designate the location of your additional data, click **Browse**.
Save output data	specifies whether to save the predictions and diagnostic statistics in an output data set: **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Additional statistics	specifies whether to save any additional statistics. ■ **Residuals** saves the likelihood residual that is used to identify poorly fitted observations, the standardized Pearson (chi) residual that is used to identify observations that are poorly accounted for by the model, and the standardized deviance residual that is used to identify poorly fitted observations. ■ **Prediction limits** saves the upper and lower confidence limits for the predicted value.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created

- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

45

Linear Regression

About the Linear Regression Task

Linear regression analysis attempts to assign a linear function to your data by using the least squares method. Using the Linear Regression task, you can perform linear regression analysis on multiple dependent and independent variables.

For example, you might use regression analysis to find out how well you can predict a child's weight if you know the child's height. Suppose that a SAS data set contains the height and weight measurements of 19 children. By using weight as the dependent variable and height as the independent variable, you can perform a linear regression analysis on this data.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	REG
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Linear Regression: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variables	specifies the variables to use as the dependent (response) variables. You must assign at least one variable to this role.
Explanatory variables	specifies the variables to use as the regressor (explanatory) variables. If you do not specify any regressor variables, each dependent variable is fit to a mean-only model. That is, each dependent variable is fit to a model with only a constant term.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analysis by a variable that you selected as a response variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable that contains the values that are relative weights for a weighted least squares fit. The values of the relative weight variable must be nonnegative. You can assign a maximum of one variable to this role.

Linear Regression: Setting the Model Options

In the selection pane, click **Model** to access these options.

Option Name	Description
Model selection method	Select a method from the drop-down list of model selection methods. If you do not select a method, the full model is used. Depending on the model selection method, additional options for that method become available. **Full model fitted (no selection)** This method is the default and provides no model selection capability. This is the model that was created when you assigned the dependent variables and quantitative (explanatory) variables. **Forward selection** The forward selection method begins with no variables in the model. For each of the explanatory variables, this method calculates *F* statistics that reflect the variable's contribution to the model, if it is included. The *p*-values for these *F* statistics are compared to the significance level that is specified for including a variable in the model. By default, this value is `0.5`. To change this significance level, enter the value in the **To enter the model** text box. If no *F* statistic has a significance level greater than this value, the forward selection stops. Otherwise, the forward selection method adds the variable with the largest *F* statistic to the model. The forward selection method then calculates *F* statistics again for the variables that remain outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant *F* statistic. After a variable is added to the model, it stays there. **Backward elimination** The backward elimination method begins by calculating *F* statistics for a model, including all of the explanatory variables. Then the variables are deleted from the model one by one until all the variables that remain in the model produce significant F statistics. The significance level is specified in the **To stay in the model** text box. By default, this value is `0.10`. At each step, the variable that shows the smallest contribution to the model is deleted. **Stepwise selection** The stepwise method is a modification of the forward selection method. In the stepwise method, variables that are already in the model do not necessarily stay there. As in the forward selection method, variables are added one by one to the model, and the *F* statistic for a variable to be added must be significant. The significance level is specified in the **To enter the model** text box. However, after a variable is added, the stepwise method checks all the variables that already included in the model and deletes any variable that does not produce a significant *F* statistic. The significance level is specified in the **To stay in the model** text box. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when either of these conditions is met: ■ no variable outside the model has a significant *F* statistic, and every variable in the model is significant at the significance level that is specified to stay in the model ■ the variable to be added to the model is the variable that was just deleted from it

Option Name	Description
Model selection method (continued)	**Maximum R-squared improvement** The maximum R-squared improvement method does not settle on a single model. Instead, it tries to find the best one-variable model, the best two-variable model, and so on. It is not guaranteed to find the model with the largest R^2 for each size. This method begins by finding the one-variable model that produces the highest R^2. Then another variable, the one that yields the greatest increase in R^2, is added. After the two-variable model is obtained, each variable in the model is compared to each variable not in the model. For each comparison, this method determines whether removing one variable and replacing it with the other variable increases R^2. After comparing all possible switches, this method makes the switch that produces the largest increase in R^2. Comparisons begin again, and the process continues until this method finds that no further switch could increase R^2. Thus, the resulting two-variable model is considered the best two-variable model that the method can find. Another variable is then added to the model, and the comparing-and-switching process is repeated to find the best three-variable model, and so on. The difference between the stepwise selection method and the maximum R^2 selection method is that in the maximum R^2 method, all switches are evaluated before any switch is made. In the stepwise selection method, the "worst" variable might be removed without considering what adding the "best" remaining variable might accomplish. **Minimum R-squared improvement** The minimum R-squared improvement method closely resembles the maximum R-squared improvement method, but the switch that is chosen is the one that produces the smallest increase in R^2. For a given number of variables in the model, the maximum R-squared and minimum R-squared methods usually produce the same best model, but the minimum R-squared method considers more models of each size. R-squared selection The R-squared selection method finds subsets of explanatory variables that best predict a dependent variable by linear regression in the given sample. You can specify the largest and smallest number of explanatory variables to appear in a subset and the number of subsets of each size to be selected. The R-squared method can efficiently perform all possible subset regressions and display the models in decreasing order of R^2 magnitude within each subset size. Other statistics are available for comparing subsets of different sizes. These statistics, as well as estimated regression coefficients, can be displayed or sent as output to a SAS data set. The R-squared selection method differs from the other selection methods in that the R-squared selection method always identifies the model with the largest R^2 for each number of variables that are considered. **Adjusted R-squared selection** The adjusted R-squared selection method is similar to the R-squared selection method, except that the adjusted R^2 statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted R^2 within the range of sizes. **Mallows' Cp selection** Mallows' Cp selection method is similar to the adjusted R-squared selection method, except that Mallows' Cp statistic is used as the criterion for model selection. Models are listed in ascending order of C_p.
Significance levels	For the forward selection method, specify a significance level to enter an explanatory variable in the model. For the backward elimination method, specify a significance level so that an explanatory variable stays in the model. For stepwise selection method, specify both significance levels.
Model fit statistics	For the R-squared, adjusted R-squared selection or Mallow's Cp selection, use the check boxes to select the model fit statistics to include in the results.
Effects to force into the model	For all methods except the full model fitted method, select the check boxes of the effects to force into the model. Use the arrows to reorder the effects. To remove an effect, clear its check box.
Include intercept	For all methods, clear this check box to remove the effects of the intercept from the analysis. The intercept is included by default.

Linear Regression: Setting Statistics Options

In the selection pane, click **Statistics** to access these options.

Option Name	Description
Details on estimates	sets the estimates options. You can choose from these options. ■ standardized regression coefficients. A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor. ■ the sequential sum of squares (Type 1) and the partial sum of squares (Type 2) along with parameter estimates for each term in the model. ■ the correlation matrix of the estimates. ■ the covariance matrix of the estimates. ■ confidence level for parameter estimates. The default is a 95% confidence level.
Correlations	specifies whether to include squared partial and squared semi-partial correlation coefficients. ■ The squared partial correlation coefficients are calculated as SS/(SS+SSE), where SSE is the error sum of squares. ■ The squared semi-partial correlation coefficients are calculated as SS/SST, where SST is the corrected total sum of squares. If you choose not to include the intercept in the model, then the uncorrected total sum of squares is used in the denominator. Each type of correlation uses both Type 1 and Type 2 sums of squares to calculate the correlation coefficients.
Diagnostics	specifies the diagnostic analyses to perform. ■ collinearity analysis. This option requests a detailed analysis of collinearity among the regressors. This includes eigenvalues, condition indices, and decomposition of the variances of the estimates with respect to each eigenvalue. ■ collinearity analysis without the intercept. This option requests the same analysis as the collinearity analysis option except that the intercept is not included in the analysis. ■ tolerance values for estimates. Tolerance for a variable is defined as 1-R2, where R2 is obtained from the regression of the variable on all other regressors in the model. variance inflation values. Variance inflation is the reciprocal of tolerance. ■ heteroscedasticity test. This option tests that the first and second moments of the model are correctly specified. ■ asymptotic covariance matrix. This option displays the estimated asymptotic covariance matrix of the estimates under the hypothesis of heteroscedasticity. ■ Durbin-Watson statistic. The Durbin-Watson statistic shows whether the errors have first-order autocorrelation. (This test is appropriate only for time series data.) The sample autocorrelation of the residuals is also produced.

Linear Regression: Generating Plots

In the selection pane, click **Plots** to access these options.

By default, all appropriate plots for the current data selection are included in the output. However, you can choose which plots to include in the output by selecting the **Custom lists of plots** option. You can choose from these options:

Plot Name	Description
Histogram plot of the residuals	creates a histogram of the fit residuals.
Residuals by predicted values plot	creates a plot of the residual versus predicted values. A residual is the difference between the observed value and the predicted value.
Studentized residuals by predicted values plot	creates a plot of the Studentized residuals versus predicted values.
Observed by Predicted values plot	creates a plot of the observed versus predicted values.
Plot Cook's D statistic	creates a plot of Cook's D statistic versus the observation number.
Studentized residuals by leverage plot	creates a plot of the studentized residuals versus the leverage. A studentized residual is the value of the residual divided by the standard error with the current observation deleted. The leverage for an observation is the value of the corresponding diagonal element in the *H* matrix.
Normal quantile plot of the residuals	creates a Q-Q plot.
Residual-Fit plot	creates side-by-side plots of the quantiles of centered fit and the residuals.
Box plot of residuals	creates a box plot of the residuals.
Diagnostic plots	creates plots of the fit diagnostics.
DFFITS plots	creates plots of DFFITS (Difference in Fit, Standardized). The DFFITS statistic is the standard influence of the observation on the predicted value. A large value indicates that the observation is very influential.
DFBETS plots	creates plots of the standardized difference in the parameter estimate due to deleting the observation. It is useful to assess the effect of an individual observation on each estimated parameter of the fitted model.
Residual plots	creates plots of the residuals versus the regressors.
Scatter plot with regression line	creates a plot of the regression line overlaid on a scatter plot of the data.

Linear Regression: Setting Prediction Options

In the selection pane, click **Predictions** to access these options.

Option Name	Description
Data to predict	identifies the data sources that you want to use to predict the output data. If you select the **Additional data** check box, this data is included.
Additional statistics	specifies the additional statistics to include. ■ Selecting **Residuals** includes the ordinary residuals, standardized residuals, and studentized residuals in the output. An ordinary residual is the difference between the observed value and the predicted value. A standardized residual is the value of the residual divided by the standard error. A studentized residual is the value of the residual divided by the standard error with the current observation deleted. ■ Selecting **Prediction limits** includes the upper and lower limits of a 100*(1-alpha)% confidence interval for each prediction and the expected mean of the dependent variable in the output.
Save output data	specifies whether to permanently save predictions and diagnostic statistics in an output data set. Here are the diagnostic statistics: ■ Cook's *D* influence statistic ■ the standard influence of observation on predicted value (the DFFITS statistic) ■ the leverage, ■ the standard error of the individual predicted value ■ the standard error of the mean predicted value ■ the standard error of the residual Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Display output and plots	displays the results. To suppress the results, clear the check box. If you clear this check box, the options on the **Plots** panel are not available. The results are displayed by default. To include the predictions in the results, select the **Show predictions** check box. This option is available only if a data source in the **Data to predict** area and the **Display output and plots** check box are selected.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

46

List Data

About the List Data Task

The List Data task prints the observations in a SAS data set, using all or some of the variables. You can create a variety of reports, ranging from a simple listing to a highly customized report that groups the data and calculates totals and subtotals for numeric variables.

For example, you can use the List Data task to create a report that sums the expenses and revenues for each sales region, compares these values to the total expenses and revenues for the company, shows the number of observations in each subgroup and in the whole report, and has a customized title that contains the name of the region.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	PRINT
Required SAS products	Base SAS
Recommended additional SAS products	None

List Data: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
List variables	prints the variables specified in the order they are specified. You must assign at least one variable to this role.
Group analysis by	specifies how to sort the table. A listing is generated for each distinct value (or BY group) in the variable or combination or variables.
Page by	prints a new page whenever the value of the specified variable changes, or when the next BY group begins. The variable that you assign to this role must also be a variable in the **Group analysis by** role. You can assign a maximum of one variable to this role.
Total of	prints the sums of the selected variable at the bottom of the listing report.
Subtotal of	prints a subtotal whenever the value of the specified variable changes or when the next BY group begins. The variable that you assign to this role must also be a variable in the **Group analysis by** role. You can assign a maximum of one variable to this role.
Identifying label	specifies how to identify the rows rather than use observation numbers.

List Data: Setting Listing Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Rows to list	specifies the rows to list. You can choose from these options: ■ **All rows** prints all of the rows. ■ **First n rows** prints the number of rows that you specify in the **Amount (n)** text box. For example, if you set *n* to 50, then the output data contains rows 1 through 50 from the input data set. ■ **First n percent of rows** prints the percentage of rows that you specify in the **Amount (n)** text box. For example, if your input data set contains 100 rows and you set *n* to 25%, then the output data contains rows 1 through 25 from the input data set. ■ **Every nth row** prints every *n*th row in the data set. You specify the value of *n* in the **Amount (n)** text box. For example, if you set *n* to 10, then the output data contains every 10th row from the input data set.
Print the row number	includes in the output a column that lists the row number for each observation. You can specify a label for this column in the **Column heading** text box. The default name of this column is `Row number`.
Use variable labels as column headings	uses the variable label instead of the variable name as the column heading.
Print number of rows	reports the number of rows in the table at the end of the output, or the number of rows in each BY group at the end of each BY group's output.

Option Name	Description
Round values before summing	rounds each numeric value to the number of decimal places in its format, or to two decimal places if no format is specified. If this option is specified, the List Data task performs the rounding before summing the variable.
Divide pages into sections	divides each page into sections if there are more columns than will fit across the page. When you clear the **Divide page into sections** option, the List Data task puts as many observations on each page as it can. If there are more columns than will fit across a page, the columns that do not fit are put on subsequent pages. Depending on the number of rows and columns in the table, this option could result in fewer printed pages. **Note:** This option affects only text output. This option does not affect results in SAS Report, HTML, RTF, or PDF format.
Heading direction	prints column headings horizontally or vertically. Select **Default** and let SAS determine the optimal arrangement for each column.
Column width	specifies how the List Data task determines column widths: ■ **Default**: List Data task determines column widths on a per-page basis. ■ **Full**: List Data task uses format width (or default width if no format specified) for all pages. ■ **Minimum**: List Data task uses smallest possible column width on a per-page basis. ■ **Uniform**: Reads entire table to determine appropriate column widths before generating output. When this option is not selected, different pages might have different widths for the same column. ■ **Uniform by**: List Data task prints all pages uniformly within each BY group. **Note:** This option affects only text output. This option does not affect results in SAS Report, HTML, RTF, or PDF format.

Split labels at	specifies that the label should be split if the variable labels contain one of the split characters (*, !, @, #, $, %, ^, &, or +). For example, for a variable label that reads "This is*a label" and the * character selected as the split character, the column heading will read:
	`This is`
	`a label`
	You do not need to select both **Use variable label as column headings** and **Split labels at**. **Split labels at** implies that you want to use variable labels.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

47

List Report Wizard

About the List Report Wizard

The List Report wizard enables you to create detail or summary reports. A detail report contains one row for every observation selected for the report. A summary report consolidates data so that each row represents multiple observations. Both detail and summary reports can contain totals that summarize numeric data for a set of rows or for all rows. You can also create separate tables or page breaks based on the unique values of a variable.

For example, you can use the List Report wizard to create a report that lists the sales for soda and coffee at grocery stores in each region. The managers in each region are grouped into a separate table. Within each table, the individual product sales and total sales are listed for each manager.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	REPORT
Required SAS products	Base SAS
Recommended additional SAS products	None

List Report Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to define the list.

List Report Wizard: Define the List

If the data source contains fewer than 20 columns, then the List Report wizard includes all the columns in the data source. These columns appear in the **Preview** pane. If your data contains more than 20 columns, then no columns are included in the list report, and the **Preview** pane is empty.

You must specify the columns to add to the report. You can add all the columns in the data source, a specific column, or the frequency.

Click **Next** to display the totals.

List Report Wizard: Display the Totals

You can choose to display the subtotals or grand totals for any numeric variables that are not classification or across variables.

To specify totals:

1. Before you can specify the totals, you must define the list.
2. In the **Select totals** pane, select the check boxes for the variables for which you want to calculate a total. To display the totals for all the variables, select the **Select all** check box.
3. To specify the type of total, the position of the totaled value, and the label for the row that contains the total, click **Edit**. The Edit Totals dialog box appears.

   You can choose from these options:

   - **Type of totals** specifies the type of totals to display in the table. You can choose to display the grand totals and the subtotals. You must select at least one type of total to display. To display all of the totals in the table, select the **Select all** check box.
   - **Position** specifies where to display the total row in the table. You can choose to display the total before or after the rows or observations that were used to calculate the total.
   - **Label for totals** specifies the label to use for the total rows. By default, this label is Total.
   - **Label for subtotals** specifies the label to use for the subtotal rows. By default, this label is Subtotal. You can also specify whether to display the label before or after the text in this row. By default, no space is inserted between the label and any preceding or subsequent text. To include a space, you must add the space before or after the label in the **Label for subtotals** box.
4. Click **Next** to specify title and footnotes.

List Report Wizard: Specify the Titles and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1. Before you can specify the titles and footnotes, you must display the totals for any numeric variables.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

3 Click **Finish** to run the task.

48

Logistic Regression

About the Logistic Regression Task

Binary responses, such as success and failure, and ordinal responses, such as normal, mild, and severe, arise in many fields of study. Logistic regression analysis investigates the relationship between these discrete responses and a set of explanatory variables.

You might want to use logistic regression to determine the effects of a clinical study. For example, suppose that you use the Logistic Regression task to analyze the results from a study of the analgesic effects of treatments on elderly patients with neuralgia. Two test treatments and a placebo are compared. The response variable is whether the patient reported pain or not. Researchers record the age and gender of the patients and the duration of the complaint before the treatment begins. From the output from the logistic regression task, you can determine whether the pain reduction effects of these two test treatments are different.

Note: When you open the Logistic Regression task, you might notice a slight delay as SAS Enterprise Guide determines which variables have binary values. Once this analysis is complete, the task provides information about the variable that you assign to the **Dependent variable** role.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	LOGISTIC
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Logistic Regression: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variable	specifies the variable to use as the dependent (response) variable. You must assign only one variable to this role.
Quantitative variables	specifies the variables to use as the independent (explanatory) variables. You can also specify the units of change for the quantitative variables. For example, to generate estimated custom odds ratios in the results, enter `-2` in the **Units** box so that an odds ratio represents the change in the odds when the variable is decreased by two units.
Classification variables	specifies the classification variables to use in the analysis. They can be character or numeric variables. If they are numeric, the values must be discrete rather than continuous. To select the parameterization method for each variable, select the classification variable. From the **Coding style for *variable-name*** area, select whether the variable's parameterization method should be effect coding (**Effects**) or reference cell coding (**Reference**).
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analysis by a variable that you selected as a response variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies one variable to use as relative weights for the corresponding values of the variable that you selected as the response variable. If a variable value is negative or missing, it is excluded from the analysis. To normalize the weighting of the variable, select the **Normalize** check box.

Logistic Regression: Building a Model

Specifying the Response

In the selection pane under the Model heading, click **Response** to access these options.

Role Name	Description
Response type	The available response types depend on the number of levels in your response variable. ■ If your response variable has only two levels, the response type is binary. ■ If your response variable has more than two levels, you can choose either ordered or unordered as the response type.
Type of model	You can select from these model types: ■ **logit** displays the output from the log odds function, which is the default selection. The binary logit model is used when there are two response categories. The cumulative logit model is used when there are more than two response categories. ■ **probit** displays the output from the inverse standard normal probability integral function. The binary probit model is used when there are two response categories. The cumulative probit model is used when there are more than two response categories. ■ **complementary log-log** displays the output from the complementary log-log function. The binary complementary log-log model is used when there are two response categories. The cumulative complementary log-log model is used when there are more than two response categories. ■ **log-log** displays the output from the log-log function. The binary log-log model is used when there are two response categories. The cumulative log-log model is used when there are more than two response categories. ■ **glogit** displays the output from the generalized logit function. In the generalized logit model, each nonreference category is contrasted with the reference category. **Note:** This option is available only if you specified **Unordered** as the response type.
Response levels	The Response levels box lists in ascending order the formatted levels for the response variable.
Fit model to level	You can specify how to fit the model to a level. ■ If you selected **Binary** as the response type, then from the Fit model to level drop-down list select the event category for the binary response model. ■ If you selected **Ordered** as the response type, then you can use the values in the **Fit model to level** drop-down list to specify the order of the response categories. If you select the first ordered value, then the response categories are ordered as they appear in the **Response levels** box. If you select the last ordered value, then the response categories are in reverse order.
Reference level	This option is available only if you selected **Unordered** as the response type. For the generalized logit model, you can specify the reference category from the **Reference** level drop-down list. For the generalized logit model, each logit contrasts a nonreference category with the reference category.

Model Options

About Models

In the selection pane, click **Model** to access these options.

By default, no effects are specified, which results in the task fitting an intercept-only model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects.

Specify Main Effects

1 Select the variable name in the **Class and quantitative variables** box.

2 Click **Main** to add the variable to the **Effects** box.

Specify Crossed Effects (Interactions)

1 Select one or more variable names. To select more than one variable, press Ctrl while you select the additional variable names.

2 Click **Cross**. If you select only one variable, then the variable is crossed with itself.

Specify Nested Effects

1 Select one or more main effects or crossed effects that you want to nest.

2 Select one or more Class variables within which you want to nest the main effects or crossed effects.

3 Click **Nest**. The nested effect appears in the **Effects** box in the form *variable 1*(*variable 2*).

Create Multiple Crossings of Two or More Selected Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Factorial** to add the factorial effects to the **Effects** box. By default, each combination of two variables is used.

Create Polynomial Crossing of Quantitative Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Polynomial** to add the polynomial effects to the **Effects** box.

Remove the Effects of the Intercept from the Analysis

Clear the **Include intercept check** box. If you clear this check box, then you must have specified at least one effect. By default, the intercept is included.

Specifying a Model Selection Method

In the selection pane under the Model heading, click **Selection** to access these options.

Option Name	Description
Model selection method	Select a method from the drop-down list of model selection methods. If you do not select a method, the full model is used. Depending on the model selection method, additional options for that method become available. Here are the available methods: ■ **Full model fitted (no selection)** is the default and provides no effect selection. This is the model that was created when you assigned the dependent variables and quantitative variables is used to fit the model. ■ **Forward selection**begins with no effects in the model. For each of the explanatory variables, this method calculates *F* statistics that reflect the variable's contribution to the model, if it is included. The *p*-values for these *F* statistics are compared to the significance level that is specified for including a variable in the model. By default, this value is `0.05`. To change this significance level, enter the value in the **To enter the model** text box. If no *F* statistic has a significance level greater than this value, the forward selection stops. Otherwise, the forward selection method adds the variable with the largest *F* statistic to the model. The forward selection method then calculates *F* statistics again for the variables that remain outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant *F* statistic. After a variable is added to the model, it stays there. ■ **Backward elimination** begins by calculating *F* statistics for a model, including all the explanatory variables. Then the variables are deleted from the model one by one until all the variables that remain in the model produce significant *F* statistics. The significance level is specified in the To stay in the model text box. By default, this value is `0.05`. At each step, the variable that shows the smallest contribution to the model is deleted. ■ **Stepwise selection** is a modification of the forward selection method. In the stepwise method, variables that are already in the model do not necessarily stay there. As in the forward selection method, variables are added one by one to the model if the *F* statistic is significant. The significance level is specified in the To enter the model text box. However, after a variable is added, the stepwise method checks all the variables that are already included in the model and deletes any variable that does not produce a significant *F* statistic. The significance level is specified in the **To stay in the model** text box. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when either of these conditions is met: □ no variable outside the model has a significant *F* statistic, and every variable in the model is significant at the significance level that is specified to stay in the model. □ the variable to be added to the model is the variable that was just deleted from it. ■ **Best subset** uses an algorithm to find a specified number of models with the highest chi-square statistic for all possible model sizes, from 1, 2, 3 effect models, and so on, up to the single model containing all of the quantitative variables.

Option Name	Description
Significance levels	■ For the forward selection method, specify a significance level to enter a quantitative (explanatory) variable in the model. ■ For the backward elimination method, specify a significance level so that a quantitative variable stays in the model. ■ For the stepwise selection method, specify both significance levels.
Effects to force into the model	Select the effect that you want to add to the model. The variable is added to the **Effects to force into the model** box. You can reorder these variables by selecting the variable and clicking the arrow buttons. Note: This option is not available if you select the **Full model fitted (no selection)** model selection method.

Logistic Regression: Setting the Model Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Details on estimates	specifies whether to display the correlation matrix of the parameter estimates, the covariance matrix of the parameter estimates, or both.
Model fit assessment	specifies these model fit statistics: ■ **Influence statistics** displays diagnostic measures for identifying influential observations in the case of a binary response model. ■ **Hosmer and Lemeshow goodness-of-fit test** displays the Hosmer and Lemeshow goodness-of-fit test for the case of a binary response model. ■ **Deviance and Pearson goodness-of-fit tests** displays the Pearson chi-square statistic, the deviance, their degrees of freedom, the ratio of each statistic divided by its degrees of freedom, and the corresponding *p*-value. ■ **Generalized R-squared** displays a generalized coefficient of determination for the fitted model. Note: The **Influence statistics**, the **Hosmer and Lemeshow goodness-of-fit test**, and the **Deviance and Pearson goodness-of-fit tests** options are available only if you assign a binary variable to the **Dependent variable** role.
Show classification table	classifies the input binary response observations according to whether the predicted event probabilities are above or below some cutpoint value in the range (0,1). An observation is predicted as an event if the predicted event probability exceeds the cutpoint value. The default table displays the classification for a range of probabilities from the smallest estimated probability (rounded down to the nearest 0.02) to the highest estimated probability (rounded up to the nearest 0.02) with 0.02 increments. To supply cutpoints other than the default list, enter one or more values in the **Critical probability values (cutpoints)** text box. Each value must be between 0 and 1 inclusive. Note: This option is available if the variable that you assigned to the **Dependent variable** role is binary.

Option Name	Description
Confidence limits	specifies the confidence intervals for the parameters in the **Parameters** area and the odds ratio in the **Conditional odds ratios** area: ■ **Wald** computes confidence intervals based on individual Wald tests. ■ **Profile likelihood** computes confidence intervals based on the profile likelihood. The construction of this interval is derived from the asymptotic chi-square distribution of the generalized likelihood ratio test. You can also select a value from the **Confidence level** drop-down list to set the level of significance for the regression parameters or the odds ratios.
Model fitting methods	specifies how to fit the model. By default, the task automatically determines the best-fitting method for the model. From the **Fitting technique** drop-down list, you can choose the **Fisher scoring** or **Newton-Raphson** optimization technique for estimating the regression parameters. Note: If you selected **Unordered** as the response type, then the Fisher scoring optimization technique is not available. Select the **Firth's penalized likelihood** check box to perform Firth's penalized maximum likelihood estimation, which reduces bias in the parameter estimates. This method is useful in cases of separability, as often occurs when the event is rare, and is an alternative to performing an exact logistic regression. Note Note: The **Firth's penalized likelihood** check box is available only if you assign a binary variable to the **Dependent variable** role.

Logistic Regression: Generating Plots

In the selection pane, click **Plots** to access these options.

By default, all appropriate plots for the current data selection are included in the output. However, you can choose which plots to include in the output by selecting the **Custom lists of plots** option. You can choose from these options:

Plot Name	Description
Odds ratio	plots generates plots of the profile likelihood confidence limits for odds ratios and the Wald's confidence limits for odds ratios. This option is available only if you selected a **Conditional odds ratios** option on the Model > Options panel.
ROC plots	generates a plot of the receiver operating characteristic. The area under the ROC curve is determined by the statistic *c* in the "Association of Predicted Probabilities and Observed Responses" table.

Plot Name	Description
Influence plots	displays the following plots: ■ index plots of Pearson (chi-square) residual for identifying observations that are poorly accounted for by the model ■ the deviance residual for identifying poorly fitted observations ■ leverage ■ the confidence interval displacement diagnostic that measures the influence of individual observations on the regression estimates ■ the confidence interval displacement diagnostic that measures the overall change in the global regression estimates due to deleting the individual observation ■ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation ■ the change in the deviance that is attributable to deleting the individual observation
DfBETAS plots	displays plots of the standardized difference in the parameter estimate due to deleting the observation. It is useful to assess the effect of an individual observation on each estimated parameter of the fitted model.
Phat plots	displays these plots: ■ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation ◦the change in the deviance that is attributable to deleting the individual observation ■ confidence interval displacement ■ the leverage versus the predicted event probability
Leverage plots	displays these plots: ■ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation ■ the change in the deviance that is attributable to deleting the individual observation ◦confidence interval displacement ■ the predicted probability versus the leverage
DPC plots	displays these plots: ■ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation versus the predicted event probability ■ the change in the deviance that is attributable to deleting the individual observation versus the predicted event probability In these plots, the data markers are colored according to the value of the confidence interval displacement.
Effect plots	displays plots of the predicted probabilities for a value of the dependent variable versus the model effects.

Note: The ROC plots, Influence plots, DfBETAS plots, Leverage plots, Phat plots, and DPC plots are available only if you assigned a binary variable to the Dependent variable role.

Logistic Regression: Setting Prediction Options

In the selection pane, click **Predictions** to access these options.

Option Name	Description
Data to predict	identifies the data sources that you want to use to predict the output data. If you select the **Additional data** check box, this data is included.
Save output data	specifies whether to permanently save predictions and diagnostic statistics in an output data set. ■ **Predictions** saves the predicted probability of each response level. For a response variable Y with three levels, 1, 2, and 3, the individual probabilities are Pr(Y=1), Pr(Y=2), and Pr(Y=3). • ■ **Diagnostic statistics** saves the DFBETAS diagnostics statistic, the confidence interval displacement diagnostics (C and CBAR), the *H* matrix, and the estimates of the linear predictor with standard error estimates. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Additional statistics	■ Select **Residuals** to save these statistics:, , the change in deviance □ deviance residual □ Pearson (chi-square) residual □ the change in the deviance that is attributable to deleting the individual observation □ the change in the chi-square goodness-of-fit statistic that is attributable to deleting the individual observation ■ Select **Prediction limits** to include the upper and lower limits of a confidence interval for each prediction and the expected mean of the dependent variable in the output.
Display output and plots	displays the results. To suppress the results, clear the check box. If you clear this check box, the options on the **Plots** panel are not available. The results are displayed by default. To include the predictions in the results, select the **Show predictions** check box. This option is available only if a data source in the **Data to predict** area and the **Display output and plots** check box are selected.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

49

Map Chart

About the Map Chart Task

The Map Chart task creates a two-dimensional (choropleth) or three-dimensional (block and prism) color map that shows the variation in the value of a response variable for different geographical areas, such as counties, states, and countries.

This example shows the differences between each type of map. On a map of the United States, you want to show the number of hazardous waste sites in each state.

- In a choropleth map, the number of sites in each state (the response variable) is represented by a different shade of red. States with the largest number of hazardous waste sites are in dark red. States that do not contain any hazardous waste sites are white.
- In a block map, the number of sites in each state is represented by the height and color of the block that appears in the center of each map area.
- In a prism map, the number of sites in each state is represented by the height and color of the polyhedrons (raised polygons) for each map area.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GMAP

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Map Chart: Selecting Type of Map

In the selection pane, click **Map Chart** to access these options.

Role Name	Description
2D Cholorpleth Map	creates a two-dimensional choropleth map in which values of the specified response variables are represented by varying colors. To quickly view your map data set, select the **Create a blank map** check box. When you generate a blank map, you do not specify a response data set, so the result is a simple choropleth map.
Riser Map	creates a three-dimensional block map in which levels of magnitude of the specified response variables are represented by blocks of varying height and color.
3D Prism Map	creates a three-dimensional prism map in which levels of magnitude of the specified response variables are represented by polyhedrons (raised polygons) of varying height and color.

Map Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

The Map Chart task requires two input data sets:

- the response data source which contains the information that is overlaid on the map. Examples of response data could be sales figures or population demographics. By default, the data source that you selected before opening the task is the response data source for the task.
- the map data source that contains the X and Y coordinates for the map. After you open the task, you must specify a map data source.

The name of each input data source appears in the **Map Data Source** and **Response Data Source** fields. If you are creating a blank map, then the map data source and the response data source are the same.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Id	identifies the variables in the map and response data sets that define the map area. Every column that you assign to the ID variable role must appear in both the map and response data sets. ID variables can be numeric or character and should have the same name, type, and length in both the map and response data sets.

Role Name	Description
Response	specifies the variable in the response data set which contains the response values that are to be represented on the map. How the values of the response variable are displayed depends on the type of map. You can assign only one response variable. ■ In choropleth maps, the values of the response variables are represented by different colors that fill the map areas. ■ In block maps, the values of the response variables are represented by varying heights, patterns, or colors. ■ In prism maps, the values of the response variables are represented by raised map areas.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value.

Map Chart: Setting Appearance Options

Setting Map Options

In the selection pane under the Appearance heading, click **Mapping** to access these options.

Role Name	Description
Including all map areas from the map data set	specifies whether to include in the output any geographic subdivisions that contain no data.
Generate a separate repsonse level for each different value of the response variable	specifies whether to use a different color for each unique value of the response variable.
Accept missing values as valid in the response variable	specifies whether to include missing values as valid levels in the response variable.
Collect response values as percentage of the whole	specifies whether to collect all response values (or their statistic) and chart each region as a percentage of the whole. Select the **Specify the statistic to be graphed** option if you want to specify the statistic that is used to calculate this percentage. If you do not specify a statistic to be graphed, then the response value for only the first observation in each region is counted.
Specify the statistic to be graphed	specifies the statistic to calculate for the response variable. By default, the response value of only the first observation in each region is counted. However, if your response variable contains character values, you can choose to graph the first observation or the frequency. If your response variable contains numeric values, you can choose to graph the first observation, sum, frequency, or mean values of the response variable.
Specify the number of response levels to be graphed	specifies the number of colors that are used to represent the mapped data values.

Role Name	Description
Specify the density of map observations that are used	specifies how close a map point is to other map points. If there are many map points in close proximity (high density), then it is possible to eliminate a number of them without seriously degrading the quality of the map. You can specify an integer from 0 to 6 for this option. If you do not specify a value, then all the observations in a map data set are displayed. **Note:** This option is available only if your map data set contains a variable called DENSITY.

Specifying the Layout

In the selection pane under the Appearance heading, click **Layout** to access these options.

For all types of maps, you can specify colors for these areas on the map:

- the fill for the geographic areas that do not have a value ("empty values")
- the outline for the geographic areas that do not have a value
- the outline for the geographic areas that do have a value

You can also specify the width of the outline.

If you are creating a block map, then you can also specify the shape of the block to use on the map, the color of the outline for the blocks, and the size (or width) of the blocks.

Customizing the Colors

In the selection pane under the Appearance heading, click **Colors** to access these options.

The Map Chart task uses a color scheme that is shipped with SAS. However, you can create a set of custom colors by selecting the **Use custom colors** check box and selecting the colors to use. The task uses the colors in the order in which they appear in the list of custom colors.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.

Role Name	Description
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Map Chart: Previewing the Results

In the selection pane, click **Preview** to see a preview of the results.

From this preview of the results, you can decide whether you need to make any additional changes before generating the results. Because map data sets can be large, using this preview could help save you time. If you change any options in the task, click **Refresh** to generate a new preview of the results. If an error occurs when generating the results, then the SAS log appears.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

50

Mean and Range Chart

About the Mean and Range Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The Mean and Range Chart task creates mean and range charts for the subgroup means and the subgroup ranges. These charts are useful for analyzing the central tendency and the variability of a process.

You might want to use this task to analyze your quality control. For example, suppose that in the manufacture of silicon wafers, batches of five wafers are sampled, and their diameters are measured in millimeters. The measurements for 25 batches are stored in a SAS data set, which is used to create the mean and range charts. Each point on the mean chart represents the average (mean) of the measurements for a particular batch. Each point on the range chart represents the range of the measurements for a particular batch. If all the points fall within the control limits, you can conclude that the process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Mean and Range Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Process measurement	specifies the process variables to be analyzed. These variables contain the individual measurements. You must assign at least one variable to this role.
Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A measurement identifier is required, and you assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. Note: To prevent the input data set from being sorted, clear the **Sort by variables** check box. Note: You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.

Option Name	Description
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

51

Mean and Standard Deviation Chart

About the Mean and Standard Deviation Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The Mean and Standard Deviation Chart task creates mean and standard deviation charts for the subgroup means and the subgroup standard deviations. These charts are useful for analyzing the central tendency and the variability of a process.

You might want to use this task to find the distribution of the output and to determine whether a process is in statistical control. For example, suppose that a petroleum company uses a turbine to heat water into steam that is pumped into the ground to make oil less viscous and easier to extract. This process occurs 20 times daily, and the amount of power (in kilowatts) that is used to heat the water to the desired temperature is recorded. Each point on the mean chart represents the mean of the measurements for a particular day. Each point on the standard deviation chart represents the standard deviation of the measurements for a particular day. If all the points lie within the control limits, it can be concluded that the process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Mean and Standard Deviation Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Process measurement	specifies the process variables to be analyzed. These variables contain the individual measurements. You must assign at least one variable to this role.
Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A measurement identifier is required, and you assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. Note: To prevent the input data set from being sorted, clear the **Sort by variables** check box. Note: You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.

Option Name	Description
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.

Option Name	Description
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

52

Mixed Models

About the Mixed Models Task

A mixed linear model is a generalization of the standard linear model, the generalization being that the data is permitted to exhibit correlation and nonconstant variability. The mixed linear model provides you with the flexibility of modeling not only the means of your data but its variances and covariances as well.

You might use this task to model an interaction between two main effects. For example, suppose that you have measured the height of 18 individuals, who are classified according to family and gender. Using height as the dependent variable, gender as a fixed classification effect, and family as a random classification effect, you can use the Mixed Models task to fix a mixed linear model to this data.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	MIXED
Required SAS products	Base SAS, SAS/STAT

Requirement Name	Procedure and Product Names
Recommended additional SAS products	SAS/GRAPH

Mixed Models: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variable	specifies the variable to use as the dependent (response) variable. You must assign one variable to this role.
Quantitative variable	specifies the variables to use as the continuous independent (explanatory) variables.
Classification variables	specifies the variables to use as the discrete independent effects. Variables that you assign to this role can be numeric or character, but they should have a limited number of discrete values.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.
Relative weight	specifies the variable that is used to weight the corresponding values of the variable that is selected as the response variable. You can assign a maximum of one variable to this role.
Identifying label	specifies the additional variables to be included in the predicted values tables.

Mixed Models: Building a Fixed Effects Model

In the selection pane, click **Fixed Effects Model** to access these options.

By default, no effects are specified, which results in the task fitting an intercept-only model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects.

Specify Main Effects

1 Select the variable name in the **Class and quantitative variables** box.

2 Click **Main** to add the variable to the **Effects** box.

Specify Crossed Effects (Interactions)

1 Select one or more variable names. To select more than one variable, press Ctrl while you select the additional variable names.

2 Click **Cross**. If you select only one variable, then the variable is crossed with itself.

Specify Nested Effects

1 Select one or more main effects or crossed effects that you want to nest.

2 Select one or more Class variables within which you want to nest the main effects or crossed effects.

3 Click **Nest**. The nested effect appears in the **Effects** box in the form *variable 1*(*variable 2*).

Create Multiple Crossings of Two or More Selected Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Factorial** to add the factorial effects to the **Effects** box. By default, each combination of two variables is used.

Create Polynomial Crossing of Quantitative Variables

1 Specify higher-degree crossings by adjusting the number in the **Degrees** field.

2 Click **Polynomial** to add the polynomial effects to the **Effects** box.

Remove the Effects of the Intercept from the Analysis

Clear the **Include intercept check** box. If you clear this check box, then you must have specified at least one effect. By default, the intercept is included.

Mixed Models: Setting the Fixed Effects Model Options

In the selection pane, click **Fixed Effects Model Options** to access these options.

Option Name	Description
Show tests associated with intercept	Select this option to include the results from hypothesis tests that were performed on the intercept.
Hypothesis test type	Each hypothesis test uses one or more estimable functions to test the hypothesis. Estimable functions are linear combinations of the parameters of the model. Type 3 is the default hypothesis test. You can add or remove a hypothesis test by selecting the corresponding check box.
Show parameter estimates	The value of the parameter estimate depends on the estimation method selected.
Confidence limits for parameter estimates	A confidence interval is an upper and lower bound that provides a range of reasonable values for the parameter that is being estimated with a prespecified level of confidence. You can either enter a value for the confidence level or select one from the drop-down list. The default value is a 95% confidence level.
Estimation method	An estimation method is the technique that is used to estimate the elements of the variance and covariance matrix portion of the model. The default method is **Residual maximum likelihood**. You can select a different estimation method from the drop-down list.
Degrees of freedom method	Degrees of freedom is a statistic that refers to the number of independent pieces of information that are contained in a sum. You can select one of the following methods from the drop-down list. The default method is the **Default**. When you select this option, the Mixed Model task chooses the default method depending on your model specification and whether your model contains any random and repeated effects. The following list shows how default method is selected depending on the presence of a random or repeated effect: ■ If your model contains random effects, the default method is the containment method. ■ If your model contains only repeated effects, the default method is the Between and within subject portions method. ■ If your model contains both random and repeated effects, the default method is the containment method. ■ If your model does not contain a random or repeated effect, the default method is the residual method.
Estimable functions	Select the **Type I**, **Type II**, and **Type III** check boxes to show the matrix coefficients of the Type I - Type III estimable functions for each model effect. Select the check box next to the type to show the matrix coefficients for a specific estimable function. **Note:** You can also show the estimable functions for the intercept if you select the **Show tests associated with the intercept** check box.
Use tuning parameters	**Note:** These tuning parameters are useful only when problems are ill-conditioned. To change the sensitivity in forming Type III functions, enter the new value in the **Sensitivity for Type III functions** box. Any element in the estimable function basis with an absolute value less than the specified value is set to 0. The default is `1E-08`. To change the sensitivity that is used when sweeping a matrix, enter the desired value in the **Singular check value** box. If a diagonal pivot element is less than the original diagonal element of the matrix, multiplied by the value that you specified, then the associated column is declared to be linearly dependent on previous columns, and the associated parameter is set to 0. The default value is `1E-07`.

Mixed Models: Specifying Random Effects

In the selection pane, click **Random Effects** to access these options.

1 Click **Add** to add an effect to the **Random effects to estimate** box.

2 For each effect in the list, specify options in the **Random effects and options** box. These options are available:

Option Name	Description
Random effects	Specify the random effects in your mixed model. By default, no random effects are used. You can also include an intercept as a random effect in the model. Use the **Effects Builder** dialog box to specify the random effects for your mixed model. To specify an effect, you must assign at least one variable to the **Quantitative variables** or **Classification variables** roles. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects. After you have specified the random effects, click **OK** to save your changes and to return to the Mixed Models task. For the steps on how to create a model, see "Mixed Models: Building a Fixed Effects Model" on page 318.
Confidence limits	Specify whether to show the *t*-type confidence limits for each of the random effect estimates.
Model subjects	Specify the subject identifier to include in your analysis. Complete independence is assumed across subjects. This option produces a block-diagonal structure in the *G* matrix with identical blocks. The *Z* matrix is modified to accommodate this block-diagonality. Use the Effects Builder dialog box to specify the subject identifier for your mixed model. To specify an effect, you must assign at least one variable to the **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects. The effect that you specified appears in the **Subject identifier** box. To remove this effect, click **Remove effect**. Click **OK** to save this subject identifier and to return to the Mixed Models task.
Group effect	specify the group identifier to include in the analysis. This option enables you to define an effect that specifies the heterogeneity in the covariance structure of the *G* matrix. All observations that have the same level of the group effect have the same covariance parameters. Each new level of the group effect produces a new set of covariance parameters that have the same structure as the original group. Use the Effects Builder dialog box to specify the group identifier for your mixed model. To specify an effect, you must assign at least one variable to the **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects. The effect that you specified appears in the **Group identifier** box. To remove this effect, click **Remove effect**. Click **OK** to save this group identifier and to return to the Mixed Models task.
Covariance structure	Specify the covariance structure of the *G* matrix.

Option Name	Description
Additional options	Specify whether to display the *G* matrix in the analysis and what *G* matrix options to use. These options are available: ■ **Show G matrix** displays the estimated *G* matrix. ■ **Show Cholesky root of the G matrix** displays the lower triangular Cholesky root of the estimated *G* matrix. ■ **Show inverse Cholesky root of the G matrix** displays the inverse Cholesky root of the estimated *G* matrix. ■ **Show inverse of the G matrix** displays the inverse of the estimated *G* matrix. ■ **Show correlation matrix** displays the correlation matrix that corresponds to the estimated *G* matrix. ■ **Includes estimates** displays the solution for the random effects parameters.

Mixed Models: Specifying Repeated Effects

In the selection pane, click **Repeated Effects** to access these options.

For each effect in the list, specify options in the **Repeated effects and options** box. These options are available:

Option Name	Description
Effects to use	Select the within-subjects effect to use in the analysis. The within-subjects effects are the variables that you assigned to the **Classification variables** role. You might want to specify a within-subjects effect if your data is not sorted by the appropriate repeated measure within a subject. For example, if time is your repeated effect and the data is not sorted by time within each subject, then you need to specify time as a within-subjects effect in order for your model to interpret the data correctly. By default, no within-subjects effects are used.
Model subjects	Specify the subject identifier to include in your analysis. For many repeated measures models, no repeated effect is required. If you specify a subject identifier, the repeated statement is used to specify the *R* matrix in the mixed model. The subject identifier defines the blocks in the *R* matrix. Complete independence is assumed across subjects. Use the Effects Builder dialog box to specify the subject identifier for your mixed model. To specify an effect, you must assign at least one variable to the **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects. The effect that you specified appears in the **Subject identifier** box. To remove this effect, click **Remove effect**. Click **OK** to save this subject identifier and to return to the Mixed Models task.
Covariance structure	Specify the covariance structure of the *R* matrix.

Option Name	Description
Group effect	Specify the group identifier to include in the analysis. This option enables you to define an effect that specifies the heterogeneity in the covariance structure of the *G* matrix. All observations that have the same level of the group effect have the same covariance parameters. Each new level of the group effect produces a new set of covariance parameters that have the same structure as the original group. Use the Effects Builder dialog box to specify the group identifier for your mixed model. To specify an effect, you must assign at least one variable to the **Classification variables** role. You can select combinations of variables to create crossed, nested, factorial, or polynomial effects. The effect that you specified appears in the **Group identifier** box. To remove this effect, click **Remove effect**. Click **OK** to save this group identifier and to return to the Mixed Models task.
R matrix options	Specify whether to display the *R* matrix in the analysis and what *R* matrix options to use. These options are available: ■ **Show the R matrix** displays the estimated *R* matrix. ■ **Show correlation matrix** displays the correlation matrix that corresponds to the estimated *R* matrix. ■ **Show inverse of the R matrix** displays the inverse of the estimated *R* matrix. ■ **Show Cholesky root of the R matrix** displays the lower triangular Cholesky root of the estimated *R* matrix.

Mixed Models: Specifying Least Squares Means

In the selection pane, click **Least Squares Post Hoc Tests** to access these options.

1 Click **Add** to add an effect to the **Effects to estimate** box.

2 For each effect in the list, specify options in the **Least squares mean test and options** box. These options are available:

Option Name	Description
Effects to use	Select the class effects to use in the analysis. The class effects are the variables that you assigned to the **Classification variables** role. By default, no class effects are used.

Option Name	Description
Comparisons	Specify the comparison method to use in the analysis. ■ **Show p-values for differences** displays the differences of the least squares means. These options are available: □ **None** does not display the differences. □ **Default** displays all of the pairwise differences. □ **All pairwise differences** displays all of the pairwise differences. This is the default. □ **Control using the first level** displays the differences with a control, which is the first level of each of the specified least squares means effects. Two-tailed tests and confidence limits are associated with these types of differences. □ **Test for less than control** tests whether the noncontrol levels are significantly smaller than the control. You can use this test for one-tailed results. □ **Test for greater than control** tests whether the noncontrol levels are significantly larger than the control. You can use this test for one-tailed results. ■ **Adjustment method for comparison** enables you to perform multiple comparison adjustments on the *p*-values and the confidence limits for differences of the least squares means.
Confidence limits	Specify whether to include the *t*-type confidence limits that are constructed for the least squares means.
Covariates	Specify if you want to include the covariates in the computation and whether to display the *L* matrix coefficients. ■ **Covariate values** specify how to calculate the covariate values. These options are available: □ **Mean values** sets all covariate effects equal to their mean values for computation of standard least squares means. □ **Mean values and adjusted cross products** sets the covariates effects equal to their mean values and incorporates an adjustment to the cross products of the covariates. □ **Per variable** enables you to specify arbitrary values to the covariates. Additional columns in the output indicate the values of the covariates. ■ **L matrix coefficients displayed** displays the *L* matrix coefficients for all least squares means effects.

Mixed Models: Generating Plots

In the selection pane, click **Plots** to access these options.

By default, all appropriate plots for the current data selection are included in the output. However, you can choose which plots to include in the output by selecting the **Custom lists of plots** option. You can choose from the following options:

Option Name	Description
Box Plot	requests box plots for the effects in your model that consist of classification effects only. Note that these effects can involve more than one classification variable (interaction and nested effects), but they cannot contain any continuous variables. By default, the box plots are created based on the raw residuals for the qualifying effects in the fixed effects model, the random effects model, and the repeated effects model.
Distance plot	creates a plot of the likelihood or restricted likelihood distance.
Influence estimate plots	creates panels of the delete estimates in an influence analysis.
Influence statistics panel	creates panels of influence statistics. For iterative influence analysis, the panel shows the Cook's *D* and CovRatio statistics for fixed effects and covariance parameters, enabling you to gauge the impact on estimates and precision for both types of estimates. In noniterative analysis, only statistics for the fixed effects are plotted.
Raw residual plot	creates a panel of raw residuals. By default, the conditional residuals are produced.
Studentized residual plot	creates a panel of studentized residuals. By default, the conditional residuals are produced.
Pearson residual plot	creates a panel of Pearson residuals. By default, the conditional residuals are produced.
PRESS residual plot	creates a plot of PRESS residuals or PRESS statistics. These are based on a "leave-one-out" or "leave-set-out" prediction of the marginal mean.
Scaled residual plot	creates a panel of residual graphics based on the scaled residuals--the response and marginal residuals that have been scaled by the inverse Cholesky root of the marginal variance-covariance matrix.

Mixed Models: Setting Predictions Options

In the selection pane, click **Predictions** to access these options.

Option Name	Description
Data to predict	identifies the data sources that you want to use to predict the output data. If you select the **Additional data** check box, designate the location of your additional data, by clicking **Browse**.
Save output data	specifies whether to save the predictions and predicted means in an output data set: **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Show predictions	shows the predictions in the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

53

Model Scoring

About the Model Scoring Wizard

The Model Scoring wizard enables you to score a data set against an existing SAS Enterprise Miner predictive model. Ranking or scoring raw data sets is the objective of most data mining problems. You can use a predictive model to score data sets that contain, for example, a list of loan applicants, direct marketing respondents, or potential instances of fraud. You can also use this task to create one or more segment variables for the input data set.

Before you can access a model from the Model Scoring wizard, the model that creates the scoring code must first be created and validated in SAS Enterprise Miner. The scoring code that was generated by the model must also be saved in the SAS Metadata Repository.

You might use this wizard to evaluate data tables of customer attributes. From this analysis, you can create a mailing list of customers who are most likely to make a purchase. You could use the score code from another model to assign each customer to a market segment, based on one or more attribute scores.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	SAS Enterprise Miner
Recommended additional SAS products	None

Model Scoring: Select the Data

By default, the Model Scoring wizard uses the active data in the project.

Click **Next** to select a scoring model.

Model Scoring: Select a Scoring Model

1 Before you can select a scoring model, you must select the data.

2 Click **Browse** to select a scoring model. The Model Scoring window appears. On the **Browse** tab, select the model that you want to use.

   After you have selected the model, click **OK**. The directory path for the model appears in the **Scoring Model** box and its properties appear in the **Model Attributes** box. To view the SAS code that is generated by the model, click **Source**.

3 Click **Next** to map the data variables to the required model inputs.

Model Scoring: Map Data Variables to Required Model Inputs

1 Before you can map the data variables to the required model inputs, you must select a scoring model.

2 For each required model input, select the data variable to use from the source data. By default, model inputs and data variables that have the same name and type are mapped. To change the data variable that is mapped to a model input, select the variable and click **Edit**. From the drop-down list that appears, select the data variable to use.

3 Click **Next** to specify the output columns.

   Note: The **Next** button is available only after the data variables have been mapped to all the required inputs.

Model Scoring: Select the Output Data Columns

1 Before you can specify the output data columns, you must map the data variables to the required inputs.

2 From the **Data Variables** tab and **Model Outputs** tab, select the items to include in the output. The **Data Variables** tab includes all the variables from the input data set. The **Model Outputs** tab includes **Date**, **DateTime**, and **ModelKey**. From each tab, select the items that you want to include in the output and drag them to the **Selected Columns** pane. The variables and model outputs in the **Selected Columns** pane will appear in the output. Use the up and down arrows to specify the order of the variables and outputs.

3 Click **Next** to save the output data.

Model Scoring: Save the Output Data

Before you can save the output data, you must select the output columns.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Click **Next** to confirm your selections.

The output data set contains the original data plus the scores, such as the predicted values, residuals, and classification results. For information about the variables in the output data set, see Interpreting the results.

Model Scoring: Confirm Your Selections

Before you can confirm your selections, you must select the output data columns.

The last step contains a summary of your selections. To change any of these selections, click **Back**. To run the task, click **Finish**.

Model Scoring: Interpreting the Results

The output data set contains the original data plus the scores, such as the predicted values, residuals, and classification results. The names of the score variables are usually formed by adding a prefix to the name of the target variable, the input variable, or the input data set.

The following table lists each of these prefixes, provides a brief description of the prefix, specifies whether a target variable is required, and specifies whether this variable is in the decision matrix. For more information about interpreting your results, see the Help for SAS Enterprise Miner.

Prefix	Description	Target Needed?	Decision Matrix?
AOV_	Identifies the bin number of the interval variable that is broken into 16 equally spaced groups	No	No
B_	Bin variable	Yes	No
BL_	Best possible loss of any decisions, *–B(i)*	Yes	Yes
BP_	Best possible profit of any decisions, *B(i)*	Yes	Yes
CL_	Loss computed from the target value, *–C(i)*	Yes	Yes

Prefix	Description	Target Needed?	Decision Matrix?
CP_	Profit computed from the target value, *C(i)*	Yes	Yes
D_	Label of the decision chosen by the model	No	Yes
EL_	Expected loss for the decision chosen by the model, *–E(i)*	No	Yes
EM_CCF	Average credit cost factor	No	No
EM_CLASSIFICATION	Normalized category that the case is classified into	No	Yes
EM_DECISION	Label of the decision chosen by the model	No	Yes
EM_EVENTPROBABILITY	Posterior probability of the target event	No	No
EM_EXPOSURE	Average exposure value	No	No
EM_LGD	Average loss given the default value	No	No
EM_PD	Average predicted value	No	No
EM_PREDICTION	Predicted value	No	No
EM_PROBABILITY	Maximum posterior probability that corresponds to the level that was assigned to the EM_CLASSIFICATION variable	No	No
EM_PROFITLOSS	Expected profit or loss	No	Yes
EM_SEGMENT	Segment variable	No	No
EP_	Expected profit for the decision chosen by the model, *E(i)*	No	Yes
F_	Normalized category that the case comes from	Yes	No
G_	Identification number of the group that the observation belongs to	No	No
GRP_	Identification number of the group that the observation belongs to	No	No

Prefix	Description	Target Needed?	Decision Matrix?
I_	Normalized category that the case is classified into	No	No
IC_	Investment cost, *IC(i)*	No	Yes
IMP_	Imputed variable	No	No
LBL	Text label for the variable to use in tables and plots	No	No
M_	Indicator variable that identifies missing imputed observations or indicator variable that identifies filtered observations	—	No
_N	Neighbor variable; lists the neighbor observations	Yes	No
NODE	Numeric ID for each leaf in a decision-tree output data set	Yes	No
P_	Predicted values and posterior probabilities	No	No
R_	Plain residuals; target minus prediction	Yes	No
REPL_	Replaced variable	Yes	No
ROI_	Return on investment, *ROI(i)*	Yes	Yes
SCORECARD_BIN	Bin assigned to each observation	No	No
SCORECARD_POINTS	Total scorecard points for each individual	No	No
SEGMENT	Segment variable	No	No
SOM_DIMENSION1	Identifies rows in a self-organizing map (SOM)	No	No
SOM_DIMENSION2	Identifies rows in a SOM	No	No
SOM_ID	SOM variable	No	No
SOM_SEGMENT	Identifies clusters generated by a SOM	No	No
U_	Unformatted category that the case is classified into	No	No

Prefix	Description	Target Needed?	Decision Matrix?
WOE_	Identification number of the group that the observation belongs to	No	No

Scored output data sets also contain a variable named _WARN_ that indicates problems in computing predicted values or making decisions. _WARN_ is a character variable that either is blank, indicating that there were no problems, or contains one or more of the following character codes:

Code	Meaning
C	Missing cost variable
M	Missing inputs
P	Invalid posterior probability (for example, <0 or >1)
U	Unrecognized input category

54

Nonlinear Regression

About the Nonlinear Regression Task

The Nonlinear Regression task produces least squares or weighted least squares estimates of the parameters of a nonlinear model. Nonlinear models are more difficult to specify and estimate than linear models. Instead of simply listing regressor variables, you must select the regression expression, declare parameter names, and supply initial parameter values. Some models are difficult to fit, and there is no guarantee that the procedure can fit the model successfully.

You might use this task if you need to create a nonlinear model for your experiment. For example, suppose that you want to study the relationship between concentration and velocity for a particular pair of enzyme and substrate. You record the reaction rate (velocity) that is observed at different substrate concentrations and save these values in a SAS data set. The Nonlinear Regression task enables you to calculate the estimates for each parameter, the associated asymptotic standard error, and the asymptotic confidence interval.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	NLIN
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	None

Nonlinear Regression: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variable	specifies the variable to use as the dependent (response) variable. You must assign exactly one variable to this role.
Explanatory variable	specifies the variables to use as the independent (explanatory) variables. You must assign exactly one variable to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analysis by a variable that you already assigned to another role.

Nonlinear Regression: Setting the Model Options

In the selection pane, click **Model** to access these options.

Select a nonlinear model expression from the drop-down list. This expression is used to evaluate the predicted values.

In the **Parameters for selected model** box, specify the a, b, and c parameter types as either a single value or range of values. If you specify single value as the parameter type, then the parameter is assigned the value in the **Single value** box. If you specify range of values as the parameter type, then you can also specify the first, last, and increment values.

To set restrictions, or bounds, on the parameters, click **Parameter Bounds** and specify the bounds in the Parameter Bounds window.

Note: The Parameter Bounds window validates only the syntax of the parameter bounds that you enter. It does not verify that the bounds make mathematical sense. For example, you can enter a bound such as 2 < b < 1, even though doing so will cause the model to fail to converge.

Nonlinear Regression: Setting Computation Options

Specifying Iteration Options

In the selection pane, click **Computation** to access these options.

Plot Name	Description
Method	Select a method from the **Method** drop-down list. Available options include **Gauss-Newton**, **Marquardt**, **Newton**, and **Steepest descent (Gradient)**.
Maximum iterations	Specify the number of iterations that the task performs before it stops trying to converge. This value must be a positive integer. By default, the maximum number of iterations is `100`.
Use the Moore-Penrose inverse in parameter estimation	Select this check box to use the Moore-Penrose inverse in parameter estimation.
Replace the mean square error	Select the check box, and then specify a value to replace the mean square error for computing the standard errors of the estimates in the text box.
Minimization tuning	Specify a convergence criteria or singularity criterion for the model in the text boxes. For all iterative methods, the relative offset convergence measure of Bates and Watts is used by default to determine convergence. The default is `1.0E-5`. The singularity criterion is the absolute magnitude of the smallest pivot value that is allowed when inverting the Hessian or approximation to the Hessian. The default is `1.0E-8`.

Specifying a Step-size Search Method

In the selection pane, click **Computation** to access these options.

Plot Name	Description
Step-size search method	Select a search method. The following options are available: ■ **No step-size search** removes the restriction that the objective value must decrease at every iteration. Step halving is still used to restrain the parameter estimates within specified bounds and to ensure that the number of observations that can be evaluated does not decrease. This option is useful for iteratively reweighted least squares problems. This is the default. ■ **Step halving** searches for the smallest sum of squared errors (SSE). ■ **Cubic interpolation** estimates the step size by using a cubic interpolation. If the estimated step size does not result in a decrease in SSE, step halving is used. ■ **Golden section search** determines the step size by a golden section search. The parameter Tau determines the length of the initial interval to be searched. The parameter Rho specifies how fine the search is to be.
Maximum number of step halvings	Specify the maximum number of step halvings to perform. This value must be a positive integer.

Nonlinear Regression: Generating Plots

In the selection pane, these options appear under the **Plots** heading. Select the type of plot that you want to generate. You can create plots of each type.

Note: The **Summary of requested plots** box lists all of the plots that you have requested.

Plot Name	Description
Predicted	You can generate scatter plots of observed versus predicted values for each dependent variable and observed values of each dependent variable versus each explanatory (independent) variable. For observed versus independents plots, you can also choose to superimpose confidence limits or prediction limits on all predicted plots. ■ If you specify confidence limits, then the plot shows the confidence intervals for the mean. ■ If you specify prediction limits, the plot shows the prediction intervals for the individual responses. For both plots, alpha is .05.
Residual	You can plot ordinary, standardized, and studentized residuals of each dependent variable against predicted values of each dependent variable or values of each explanatory (independent) variable. The residuals are calculated by using the following formulas: ■ An ordinary residual is the difference between the observed value and the predicted value. ■ A standardized residual is the value of the residual divided by the standard error. ■ A studentized residual is the value of the residual divided by the standard error with the current observation deleted.
Influence	You can plot leverage against the predicted values of each dependent variable or values of each explanatory (independent) variable. The leverage for an observation is the value of the corresponding diagonal element in the *H* matrix.

Nonlinear Regression: Setting Results Options

In the selection pane, click **Results** to access these options.

You can choose to permanently save statistics, parameter estimates, or both. If you choose to save parameter estimates, you can also choose to save the final estimates if the iterative limit is exceeded. You can specify the iterative limit in the Maximum iterations option.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

You can also choose to show the iteration history, Hougaard's measure of skewness, or both in the results. Hougaard's measure of skewness enables you to assess whether a parameter is close to linear or whether it contains considerable nonlinearity.

Note: These options are not available if you select the **Suppress all displayed output** check box.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

55

Nonparametric One-Way ANOVA

About the Nonparametric One-Way ANOVA Task

The Nonparametric One-Way ANOVA task consists of nonparametric tests for location and scale differences across a one-way classification. The task also provides a standard analysis of variance on the raw data and statistics based on the empirical distribution function.

You might use this task to test the null hypothesis against an alternative hypothesis. For example, suppose that 59 female patients with rheumatoid arthritis who participated in a clinical trial are assigned to two groups, active and placebo. The response status (excellent=5, good=4, moderate=3, fair=2, poor=1) of each patient is recorded. By using the Nonparametric One-Way ANOVA task, you can test the null hypothesis that there is no difference in the patient response status between the two treatment groups against an alternative hypothesis that the patient response status differs between the two treatment groups.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	NPAR1WAY
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Nonparametric One-Way ANOVA: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variables	specifies the variables to use as the dependent (response) variables. You must assign at least one variable to this role.
Independent variable	specifies the variable to use as the independent (explanatory) variable. You must assign exactly one variable to this role. Separate analyses are performed for the subgroups defined by the variable in this list.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.

Nonparametric One-Way ANOVA: Selecting Analysis

In the selection pane, click **Analysis** to access these options.

Option Name	Description
Test scores	Select analyses to perform.
Empirical distribution function	Select the **Calculate empirical distribution function statistics (EDF)** check box if you want the empirical distribution function to be performed.
Missing values	You can choose to include missing values as a legitimate classification level by selecting the check box.

Option Name	Description
Continuity correction	For Wilcoxon scores and Siegel-Tukey scores, the task incorporates a continuity correction when computing the standardized test statistic. To not use a continuity correction for any test, select the **Suppress continuity correction** check box.

Nonparametric One-Way ANOVA: Calculating Exact p-values

In the selection pane, click **Exact p-values** to access these options.

Option Name	Description
Exact p-values	By default, the task calculates asymptotic, rather than exact, *p*-values. You can choose to calculate exact *p*-values for the various analyses by selecting the appropriate check boxes. Note: The test scores have been selected in order for the check boxes in the **Exact p-values** area to be activated. To select the test scores, click **Analysis** in the selection pane. To cancel all exact p-value calculations, click **Reset**.
Computation time limit	In the **Computation time limit** area, specify the time limit for calculating each exact *p*-value, or specify the time limit as **Unlimited**. Calculating exact *p*-values can consume a large amount of time and memory.

Nonparametric One-Way ANOVA: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Saving the output data	saves some or all of your results in an output table. Select the check boxes of the statistics that you want to include in the output. Note: The test scores have been selected in order for the **Statistics to include** check boxes to be activated. To select the test scores, click **Analysis** in the selection pane. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Option Name	Description
Suppress reports	Does not display the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

56

np Chart

About the np Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The np Chart task creates np charts for the numbers of nonconformities (defects) in the subgroup samples.

You might want to use this task to monitor the number of defects in a manufacturing process. For example, suppose that an electronics company manufactures circuits in batches of 500 and uses an np chart to monitor the number of failing circuits. Thirty batches are examined, and the failures in each batch are counted. The failure counts are stored in a SAS data set, which is used to create the np chart. Each point on the np chart represents the number of nonconforming items in a particular subgroup. For example, the value that is plotted for the first batch is 5. If all the points fall within the control limits (which is 3 sigma by default), it can be concluded that the process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

np Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Number of nonconformities	specifies the variables that contain the number of nonconformities. You must assign at least one variable to this role
Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A subgroup identifier is required, and you can assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Subgroup sample size	specifies the subgroup sample sizes as the values of a variable. The subgroup sample size is required, and you assign only one numeric variable to this role. Select the **Use numeric value for subgroup (instead of a variable)** check box when the subgroup sample size is fixed and you need to specify a constant value for the size. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). Note: This role is not required if you select the **Use numeric value for subgroup (instead of a variable)** check box on the **Options** pane.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. Note: To prevent the input data set from being sorted, clear the **Sort by variables** check box. Note: You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.

Option Name	Description
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

57

One-Way ANOVA

About the One-Way ANOVA Task

A one-way analysis of variance (ANOVA) considers one treatment factor with two or more treatment levels. The goal of the analysis is to test for differences among the means of the levels and to quantify these differences. If there are two treatment levels, then this analysis is equivalent to a *t* test that compares two group means.

You might use the ANOVA task to do any of the following:

- study the effect of bacteria on the nitrogen content of red clover plants. The factor is the bacteria strain, and it has six levels.
- analyze a randomized complete block design. For example, suppose that you are interested in whether three different types of treatment have different effects on the yield and worth of a particular crop. You believe that the experimental units are not homogeneous, so you introduce a blocking factor that allows the experimental units to be homogeneous within each block. The three different types of treatment are then randomly assigned within each block
- compare the life spans of three different brands of batteries. The factor is the brand, and it has three levels.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	ANOVA
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

One-Way ANOVA: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variables	specifies the variables to use as the dependent (response) variables. You must assign at least one variable to this role.
Independent variable	specifies the variable to use as the independent (explanatory) variable. You must assign exactly one variable to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.

One-Way ANOVA: Selecting Tests

In the selection pane, click **Tests** to access these options.

You can use a one-way analysis of variance test when the variable that is assigned to the Independent variable role has two or more distinct values. You can test your model for the assumption of equal variances.

Option Name	Description
Welch's variance-weighted ANOVA	tests the group means using a weighted variance. You can use this test if the assumption of equal variances is rejected.

Option Name	Description
Tests for Equal Variance	select one of these tests: ■ **Bartlett's test** is a modification of the normal-theory likelihood ratio test. This test computes accurate Type I error rates when the distribution of the data is normal. ■ **Brown-Forsythe test** is a variation of Levene's test. Equal variances are determined by using the absolute deviations from the group medians. Although this is a good test for determining variance differences, it can be resource intensive if your data contains several large groups. ■ **Levene's test** is considered to be the standard homogeneity of variance test. This test computes the squared residuals to determine equal variance.

One-Way ANOVA: Comparing Group Means

In the selection pane under the Means heading, click **Comparison** to access these options.

Select the multiple-comparison tests for the main effect (that is, the independent variable) that you want in the output data set. Use the **Confidence level** drop-down list to select a confidence level for all selected tests. The default is a 95% confidence level.

These tests are available:

- Bonferroni *t* test
- Tukey's studentized range test (HSD)
- Duncan's multiple-range test
- Dunnett's *t* test
- Fisher's least-significant-difference test
- Gabriel's multiple-comparison procedure
- Student-Newman-Keuls multiple range test
- Waller-Duncan k-ratio test
- Scheffe's multiple comparison procedure
- Ryan-Einot-Gabriel-Welsch multiple-range test

One-Way ANOVA: Selecting Statistics for Quantitative Variables

In the selection pane under the Means heading, click **Breakdown** to access these options.

Select the descriptive statistics to calculate and display for the dependent variables at each level of the independent variable.

Option Name	Description
Mean	the arithmetic average, calculated by adding the values of a sample variable and dividing this sum by the number of observations.
Standard deviation	a statistical measure of the variability of a group of data values. This measure, which is the most widely used measure of the dispersion of a frequency distribution, is equal to the positive square root of the variance.
Standard error	the standard deviation of the sample mean. The standard error is defined as the ratio of the sample standard deviation to the square root of the sample size.
Variance	a statistical measure of dispersion of data values. This measure is an average of the total squared dispersion between each observation and the sample mean.
Number of non-missing observations	the number of observations that do not have missing values.
Number of missing observations	the number of observations for which no value is entered.
Minimum	the lowest value for an observation.
Maximum	the largest value for an observation.

One-Way ANOVA: Generating Plots

In the selection pane, click **Plots** to access these options.

Use the check boxes to generate a box plot, a means plot, or both.

Plot Name	Description
Box and whisker	A box plot displays the mean, quartiles, and minimum and maximum observations for a group. The plot represents these statistics: ■ The endpoint of the upper whisker represents the maximum value. ■ The top of the box represents the third quartile (or 75th percentile). ■ The length of the box represents the interquartile range (the distance between the third and first quartiles). ■ The dot in the box represents the mean. ■ The horizontal line in the box represents the median (or 50th percentile). ■ The bottom of the box represents the first quartile (or 25th percentile). ■ The endpoint of the lower whisker represents the minimum value.

Plot Name	Description
Means	A means plot displays the mean and standard error for the independent variable. Each point on the graph is surrounded by a variance bar. If you choose **Means**, then you can also select these options: ■ Choose whether to use standard error of the mean or standard deviation to draw the bars. The standard error of the mean is the standard deviation of the mean. The standard deviation is the square root of the sample variance. ■ Choose whether to use pooled variance when calculating the bar lengths. Pooled variance is the combined variance of two independent samples where the populations are equal. ■ Choose a height for the bars: 1, 2, or 3 standard units. ■ Choose whether to start the vertical axis at 0.

One-Way ANOVA: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Save statistics output data	saves the results to an output data set. The output data set contains sums of squares, degrees of freedom, *F* statistics, and probability levels for each effect in the model. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress all displayed output	Does not display the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

58

One-Way Frequencies

About the One-Way Frequencies Task

You can use the One-Way Frequencies task to generate frequency tables from your data. You can also use it to perform binomial and chi-square tests.

You might want to use this task to analyze the efficiency of a new drug. For example, suppose that a group of medical researchers are interested in evaluating the efficacy of a new treatment for a skin condition. Dermatologists from participating clinics are trained to conduct the study and to evaluate the condition. After the training, two dermatologists examine patients with the skin condition from a pilot study and rate the same patients. The One-Way Frequencies task can be used to evaluate the agreement of the diagnoses.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	FREQ
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

One-Way Frequencies: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the variables to be analyzed. For each variable that you assign to this role, the task creates a one-way frequency table. You must assign at least one variable to this role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already assigned to another role.

One-Way Frequencies: Setting Statistics Options

In the selection pane, click **Statistics** to access these options.

Option Name	Description
Frequency table options	Select the statistics to include in the one-way frequency table. These options are available: ■ **Frequencies and percentages with cumulatives** creates a table that contains the frequencies, cumulative frequencies, percentages of total frequencies, and cumulative percentages for each value of the analysis variable. ■ **Frequencies and cumulative frequencies** creates a table that contains the frequencies and cumulative frequencies for each value of the analysis variable. ■ **Frequencies and percentages** creates a table that contains the frequencies and percentages of total frequencies for each value of the analysis variable. ■ **Frequencies only** creates a table that contains only the frequencies for each value of the analysis variable.
Missing values	Select the **Show frequencies** check box to include missing values in the frequency tables. Select the **Include in calculations** check box to include the frequencies of missing values in binomial or chi-square tests.
Binomial proportions	Select the asymptotic test, exact *p*-values calculation, or both. For binomial proportions, specify a test proportion (null hypothesis proportion value) and confidence level.
Chi-square goodness of fit	Select the asymptotic test, exact *p*-values calculation, or both.

Option Name	Description
Exact computations	Calculating exact *p*-values can consume large amounts of memory and processing time. In the **Exact computations** area, select the **Limit computation time** check box and specify a time limit (in seconds) for the computation of each *p*-value for each crosstabulation table. The default is `900` seconds. To compute the Monte Carlo estimates of the exact p-values instead of directly computing the exact *p*-values, select the **Use Monte Carlo estimation** check box. 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.

One-Way Frequencies: Generating Plots

In the selection pane, click **Plots** to access these options.

Use the check boxes to display bar charts of the frequencies. You can display a horizontal bar chart, a vertical bar chart, or both.

One-Way Frequencies: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Saving the output data	To save the frequency tables to an output data set, select the **Create data set with frequencies and percentages** check box. To include the cumulative frequency and the cumulative percentage for one-way tables in the output data, select the **Add cumulative frequency and cumulative percentages to data set** check box. Select the analysis variables that you want to include in the data set, and then specify the location where you want to save the data set. The path of the data set must be in the form `\\server\library\member`. You can also click **Browse** to select a location. If you selected any of the binomial proportion or chi-square tests, you can permanently save the frequency tables to an output data set by selecting the **Add binomial proportion and chi-square statistic to data set** check box. Specify the location where you want to save the data set. The path of the data set must be in the form`\\server\library\member`. You can also click **Browse** to select a location. **Note:** You can permanently save the binomial proportion and chi-square statistics for only the last variable in the **Analysis variables** role.
Suppress all displayed output	Does not display the results. **Note:** In order for the task to run, you must either display the output or save the results to a data set. If you have cleared the **Create data set with frequencies and percentages** check box and selected the **Suppress all displayed output and plots** check box, then you cannot run or save the task.

Option Name	Description
Selecting the order of the output data	To specify the order of the values in the table, select the order of the output data from the **Order output data** by drop-down list. These options are available: ■ **Data set order** orders values according to their order in the input data set. ■ **Formatted values** orders values in ascending order by their formatted values. ■ **Descending frequencies** orders values by descending frequency count. ■ **Unformatted values** orders values in ascending order by their unformatted values.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

59

p Chart

About the p Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The p Chart task creates p charts for the proportions of nonconforming (defective) items in the subgroup samples.

You might want to use this task to monitor the proportion of defects in a manufacturing process. For example, suppose that an electronics company manufactures circuits in batches of 500 and uses a p chart to monitor the proportion of failing circuits. Thirty batches are examined, and the failures in each batch are counted. The failure counts are stored in a SAS data set, which is used to create the p chart. Each point on the p chart represents the proportion of nonconforming items in a particular subgroup. For example, if the number of failures in the first circuit is 5, then the value that is plotted for the first batch is 5/500 = 0.01. If all the points fall within the control limits (which is 3 sigma by default), it can be concluded that the process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

p Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Proportion nonconforming	specifies the variables that contain the number of non-conforming items. You must assign at least one variable to this role
Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A subgroup identifier is required, and you can assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Subgroup sample size	specifies the subgroup sample sizes as the values of a variable. The subgroup sample size is required, and you assign only one numeric variable to this role. Select the **Use numeric value for subgroup (instead of a variable)** check box when the subgroup sample size is fixed and you need to specify a constant value for the size. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** This role is not required if you select the **Use numeric value for subgroup (instead of a variable)** check box on the **Options** pane.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.

Option Name	Description
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.

Role Name	Description
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

60

Pareto Chart Wizard

About the Pareto Chart Wizard

Pareto charts are one of the seven basic tools for quality improvement in American and Japanese industry. A vertical Pareto chart displays the relative frequency of problems in a process as bars in decreasing order of height from left to right. A horizontal Pareto chart displays this information as bars in decreasing order of length from top to bottom.

Pareto charts are a tool for visualizing the Pareto principle, which states that a small subset of problems, the vital few, occur much more frequently than the remaining problems, the trivial many. A Pareto chart helps you establish priorities for quality improvement activities by identifying the problems that deserve the most attention and whose solutions are likely to have the greatest impact.

You might want to use the Pareto Chart task to create a comparative Pareto chart that enables you to compare the results before and after an event. For example, suppose that during the manufacture of a metal-oxide semiconductor capacitor, causes of failures are recorded before and after a tube in the diffusion furnace is cleaned. In the output, the two Pareto charts are stacked so that you can determine whether the cleaning process increased or decreased the number of defects.

Requirement Name	Procedure and Product Names
SAS procedures used	PARETO
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Pareto Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Process variable	specifies the process variable to be analyzed. This variable determines the Pareto categories for the chart. A process variable is required, and you assign only one variable to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight variable. When you assign a variable to this role, the heights of the bars represent the weighted frequencies of the categories. A relative weight variable is useful for performing a Pareto analysis that is based on cost rather than on the frequency of occurrence. You can assign only one variable to this role.
Classification variables	specifies one or two variables to use as classification variables. When you assign a variable to this role, the task creates a comparative Pareto chart that enables you to compare frequencies across classification levels.

Pareto Chart: Restricting the Number of Categories

In the selection pane, click **Options** to access these options.

Option Name	Description
Show categories	You can restrict the number of Pareto categories that the chart displays. By default, all the categories for the process are represented by bars on the Pareto chart. To modify the number of categories shown, select one of these options: ■ **N highest values** creates a chart that displays only the Pareto categories with the *n* highest values. The default value of *n* is `5`. ■ **Up to N percent of the total frequency** creates a chart that displays the Pareto categories that sum to the cumulative percentage less than or equal to *n* percent. The default value for *n* is `90`. ■ **At least N percent of the total frequency** creates a chart that displays Pareto categories whose percentages are greater than or equal to *n* percent. The default value of *n* is `5`. After you select the option, enter a value for *n* in the **N** box. To create a new category that merges all the categories that you want to exclude, select the **Display categories not included above** check box. By default, this category is labeled **Other**. You can specify a different category label that is up to 32 characters long in the **Label** box.
Include missing values for process variables	Select this check box to treat missing values as a category. This category is represented by a bar on the Pareto chart. If the process variable is a character variable, a missing value is defined as a blank internal (unformatted) value. If the process variable is numeric, a missing value is defined as any of the SAS missing values. By default, missing values are excluded from the analysis.
Save output to data set	Select this check box to save your results in an output data set. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Pareto Chart: Setting Plot Appearance Options

Setting Bar Options

In the selection pane under the Appearance heading, click **Bars** to access these options.

Option Name	Description
Bar orientation and format	Select **Vertical** to display the bars vertically. This is the default orientation. Select **Horizontal** to display the bars horizontally. Use the **Bars** drop-down list to change the bar color. Use the **Outline** drop-down list to change the bar outline color.
Label bars	You can place labels above the bars. Select **Height of bar** to label the height of each bar. Select **Cumulative percent** to label the cumulative percent of each bar. The default is to omit the label.

Option Name	Description
Show curve	By default, the Pareto chart displays the cumulative percent curve. To change the color for the curve, use the **Curve color** drop-down list. To label the curve points with their values, select the **Label curve points with values** check box. To suppress this curve and the secondary axis, clear the **Show curve** check box.
Highlight bars	Select the **Highest** check box to highlight the bars for the *n* highest values. Select **Lowest** to highlight the bars for the *n* lowest values. Use the **Color** drop-down list to change the highlight color. To highlight more than one bar, use the up and down arrows to specify a value for *n* in the **Number of bars** box. The value of *n* must be less than or equal to the number of process categories.

Pareto Chart: Setting Axis Options

In the selection pane under the Appearance heading, click **Axes** to access these options.

The orientation of the bars determines whether the axis options affect the primary vertical axis (Y) or the primary horizontal axis (X).

You can set these options:

- the axis color and width.
- the label for the primary axis.
- whether a frame is drawn around the Pareto chart and the background fill color.
- the scale for the primary axis. You can choose from these options:
  - **Percent** – The scale is the percentage of total frequency. If you assign a relative weight variable, the scale is the percentage of total weight. This is the default.
  - **Count** –The scale is in frequency units (counts). This option is not applicable if you assign a relative weight variable.
  - **Weight** – The scale is in the same units as the relative weight variable. This option is applicable only if you assign a relative weight variable.

Pareto Chart: Adding Reference Lines

In the selection pane under the Appearance heading, click **Reference Lines** to access these options.

The orientation of the bars determines whether the reference lines appear along the vertical axis or along the horizontal axis. To use reference lines, select the **Use reference lines** check box.

Pareto Chart: Adding an Inset

In the selection pane under the Appearance heading, click **Inset** to access these options.

You can enhance a Pareto chart by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative Pareto chart, an inset appears in each component chart.

Note: If you add an inset to the Pareto chart, then you should clear the **Show sample size legend** check box on the **Appearance > Options** panel. If you include an inset and the sample size legend in the Pareto chart, the boxes for the inset and the legend might overlap in the output. Therefore, it is recommended that you choose either to include an insert or display the sample size legend, but not both.

Select the **Include inset** check box to add summary statistics to the Pareto chart. Then select the statistics that you want to include in the inset. The statistics that are available depend on the options that you previously selected.

- By default, **Sample size** is the only statistic that appears in the inset.
- If you assigned a variable to the **Relative weight** role, you can select **Sum of the weights**.
- If you restricted the number of categories, these statistics are also available:
  - **Observations excluded in restricted chart** is available if there is not an **Other** category that represents the excluded categories.
  - **Number of observations in OTHER=** is available if there is an **Other** category that represents the excluded categories.
  - **Number of OTHER= categories** is available if there is an **Other** category that represents the excluded categories.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Sample size legend	Select the **Show sample size legend** check box to display a sample size legend. By default, the legend displays N=n, where *n* is the total count for the Pareto categories. If you create a comparative Pareto chart, then a legend appears in each component chart. The orientation of the bars determines the position of the legend on the chart. Note: If you add an inset to the Pareto chart, you should clear the **Show sample size legend** check box. If you include an inset and the sample size legend in the Pareto chart, then the boxes for the inset and the legend might overlap in the output. Therefore, it is recommended that you choose either to include an insert or display the sample size legend, but not both. To create a frame around the legend, select a color from the **Background** drop-down list. This color is used to fill the frame. To label the sample size legend, enter a text string of up to 32 characters in the **Label** text box.
Comparative options	If you assign at least one classification variable to an analysis role, then you can use the **Per page** boxes to change the arrangement of columns and rows in the comparative Pareto chart. If you assign one classification variable, the default arrangement is one column and two rows per page. If you assign two classification variables, the default arrangement is two columns and two rows per page. You can also use the **Label background** drop-down lists to specify the colors that are used to fill the frame area of the row labels and the column labels.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

61

Pie Chart

About the Pie Chart Task

The Pie Chart task creates simple, group, or stacked charts that represent the relative contribution of the parts to the whole by displaying data as wedge-shaped "slices" of a circle. Each slice represents a category of data. The size of a slice represents the contribution of the data to the total chart statistic. For example, a pie chart can show the sales of each store as a fraction of a chain's total sales.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GCHART
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Pie Chart: Selecting a Chart Type

In the selection pane, click **Pie Chart** to access these options.

Option Name	Description
Simple Pie	creates a chart that shows the relative contribution of the parts to the whole. The data appears as wedge-shaped "slices" of a circle. Each slice represents a category of data. The size of a slice represents the contribution of the data to the total chart statistic.
Group Pie	produces a separate pie chart for each unique value of the variable that is assigned to the **Group pies by** role.
Stacked Pie	divides the chart into concentric rings according to the values of the variables that are assigned to the **Stack pies by** role. The width of the rings, which is the same for each subgroup, is determined by the radius of the pie.
Group/Stacked Pie	produces a separate pie chart for each unique value of the variable that is assigned to the **Group pies by** role. Each chart is divided into concentric rings according to the values of the variables that are assigned to the **Stack pies by** role.

Pie Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Note: The available roles depend on the chart type that you selected.

Role Name	Description
Column to chart	The values of the column that you assign to this role determine the different slices. The column can be either character or numeric.
Group pies by	A pie chart is created for each unique value of the column that you assign to this role. All the charts are displayed on the same page.
Stack pies by	A concentric ring (or stack) is created in the pie chart for each unique value of the column that you assign to this role. This column is called a subgroup variable. **Note:** If you have selected Java as the graphics output format, then the values of the subgroup variable are not displayed as concentric rings. Instead, in the resulting graph, you drill down on a pie slice to display the values for the subgroup. However, if you have selected the Java image as the graphics output format, then the graph is a static image, and you lose the drill-down functionality that is available with the Java graphics output format.
Sum of	The column that you assign to this role determines the size of the slices.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Each pie chart is displayed on a new page.

Pie Chart: Setting Appearance Options

Setting Pie Chart Options

In the selection pane under the Appearance heading, click **Pies** to access these options.

Role Name	Description
Specifying colors for the slices	specifies the color of the slices. You can choose from these options: ■ **Default color scheme** uses the default color scheme for the slices. The default color scheme is determined by the current style. You can specify the style by using the Style Manager or the Properties dialog box for the result. In SAS Enterprise Guide, you can also specify the style for each type of result from the Options dialog box. In the SAS Add-In for Microsoft Office, you can apply styles to the SAS Report and HTML formats by using the **SAS Add-In Options** dialog box. ■ **Existing colors (Specified in a previous version)** enables you to use colors that were specified in a previous version of SAS Enterprise Guide or the **SAS Add-In for Microsoft Office**. This option is available only for SAS content that you have migrated from a previous release. ■ **Custom colors** enables you to specify the custom colors to use for the slices. Using the drop-down lists, you can specify up to 12 colors for the bar chart. If your chart needs to use more than 12 colors, then the additional colors are selected from the current default color scheme. Click **Reset** to reset the custom colors back to the default. Note: Depending on the type of donut chart, not all of the custom colors might be used.
Determining the number of slices	specifies the number of slices in the chart. By default, the number of slices is determined automatically. For continuous columns, the values are divided into ranges, with one slice for each range. Slices that would represent 4% or less of the total are combined into a slice named “Other.” To specify how many slices appear in the chart, select the **Specify number of slices** check box.

Specifying the Layout

In the selection pane under the Appearance heading, click **Layout** to access these options.

Select whether to use the name, the percentage, the statistic value, or a combination of the three in the label for each slice. For each component of the label, select one of these positions for the label:

- inside the slice
- outside the pie or donut
- outside the slice with an arrow that points to it

You can choose to make the font color for the slice labels the same color as the slices. If you want the labels to appear inside the slices, then do not request this option.

You can specify the percentage and the label for the "Other" slice. This slice is a collection of all the midpoints that have chart statistic values less than or equal to the percent-of-total. The default value of the percent-of-total is 4. Therefore, any slice that represents 4% or less of the total is put in the "Other" slice.

You can specify a text string up to 16 characters long as the label of the "Other" slice. The default label is Other. The "Other" slice is the last slice in the pie, regardless of the order of the slices.

If only one midpoint falls into the "Other" category, then the slice for that midpoint remains in its original position in the pie and retains its original label. For example, suppose a pie that shows the percent sales of beverages has these slices and percent values: Coffee 35%, Tea 15%, Soda 5%, and Milk 45%. If the percent-of-total is 5, the slice for soda remains the third slice in the pie instead of becoming the last slice, and this slice is labeled "Soda" instead of "Other."

For grouped charts, use the **Grouped Layout** area to arrange the charts. Specify the number of columns across the output area or the number of rows down the output area.

Setting Legend Options

In the selection pane under the Appearance heading, click **Legend** to access these options.

You can create legends for multiple area plots, multiple line plots, bar-line charts, contour plots, maps, stacked bar charts, pie charts, and donut charts. Legends are available for multiple measures bar charts if you assigned a variable to the **Stack** role.

For bar-line charts, you can create two legends: one for the bar chart and one for the line plot.

Legends are also available if you are creating a scatter plot, and if you assign variables to the **Vertical** and **Vertical Right** roles. Because you have multiple scatter plots on the same graph, the legend enables you to distinguish between data points for the **Vertical** variable and data points for the **Vertical Right** variable.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Setting Advanced Options

In the selection pane under the Appearance heading, click **Advanced** to access these options.

Select from the drop-down list the statistic to use when calculating the slices in the chart. If you do not assign a variable to the **Sum of** role, you can choose to calculate the frequency or percentage of each slice. By default, the frequency is calculated. If you do assign a variable to the **Sum of** role, then you can choose to calculate the sum or average of each slice. By default, the sum is calculated.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

62

Pie Chart Wizard

About the Pie Chart Wizard

The Pie Chart wizard helps you create pie charts in two or three dimensions, with options for grouped or stacked pies. Each slice of the pie represents a category of data. The size of a slice represents the contribution of the data to the total chart statistic.

Pie Chart Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to assign variables to roles.

Pie Chart Wizard: Assign Variables to Roles

1. Before you can assign variables to roles, you must select the data.
2. Assign variables to the following roles:
   - **Slice** - The values of the variable that you assign to this role determine the different slices. The variable can be either character or numeric.

     You can also specify these options:

     - **One slice for each unique data value** - For character variables, this selection has no effect. For numeric variables, one slice is drawn for each unique value.
     - **Automatically generate slices** - Numeric variables are treated as continuous, and the values are divided into ranges with one slice for each range. The number of ranges is determined automatically. Slices that would represent 4% or less of the total are combined into a slice named "Other."

- **Slice size** – The variable or statistic that you assign to this role determines the size of the slices.
  - If you assign a variable to this role, you can choose to calculate the sum or average of each slice. By default, the sum is calculated.
  - If you do not assign a variable to this role, you can choose to calculate the frequency or percentage of each slice. By default, the frequency is calculated.

3 (Optional) Assign variables to these roles:

- **Group by** - The values of the variable that you assign to this role are organized into groups. A separate pie is produced for each unique value.
- **Stack by** - The values of the variables that you assign to this role determine the number of stacked, concentric rings for each pie chart. If the graph output type is Java, then the chart has drill-down functionality rather than concentric rings.
- **Chart by** - Separate charts are generated for each BY group. The groups are determined by the values of the variable that you assign to this role. Output is generated for each BY group.

4 Click **Next** to set the appearance options.

Pie Chart Wizard: Set the Appearance Options

The Pie Chart wizard automatically determines the colors that the chart uses. You can view this color scheme from the **Schemes** drop-down list.

Specify these appearance options:

- a two-dimensional or three-dimensional chart
- a legend
- variable name on each slice on the chart
- percentage of each slice on the chart
- data value for each slice on the chart

Click **Next** to specify titles and footnotes.

Pie Chart Wizard: Specify the Titles and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1 Before you can specify the titles and footnotes, you must set the appearance options.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

3 Click **Finish** to run the task.

63

P-P Plots

About the P-P Plots Task

A process capability analysis compares the distribution of output from a process in statistical control to its specification limits in order to determine the consistency with which the specifications are met. A probability-probability plot (also referred to as a P-P plot or percent plot) is used in process capability analysis to compare the empirical cumulative distribution function (ECDF) of a variable to 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 model fits a set of measurements.

You might use a P-P plot to analyze a manufacturing process. For example, suppose that the distances between two holes that are cut into 50 steel sheets are measured and saved in a data set. The cutting process is in statistical control. You want to check whether the distances are normally distributed. You can use the P-Plot task to create a plot that is based on the normal cumulative distribution function, which enables you to determine whether or not the distances are normally distributed.

Note: Do not confuse probability-probability plots with probability plots, which compare a set of ordered measurements with percentiles from a specified distribution.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CAPABILITY

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Capability Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the numeric variables to be analyzed. You must assign at least one variable to this role. For all plots except the P-P plot, you can specify an upper limit, a target value, and a lower limit in the **Spec limits for *variable-name*** area for each analysis variable. When you specify these limits, the output automatically includes a table that shows the tests for normality, a table that shows the specification limits, and a table that shows the process capability indices.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight. When you assign a variable to this role, the value of the variable in each observation is used to compute weighted statistics. You can assign only one variable to this role. **Note:** Assigning a variable to the **Relative weight** role is not recommended when you are creating a histogram, probability plot, P-P plot, Q-Q plot, or CDF plot. If you assign a variable to this role, any output should be interpreted with caution.
Classification variables (available for the Histogram task only)	specifies one or two variables to use as classification variables. When you assign a variable to this role, the task creates a comparative histogram that enables you to compare the distribution of an analysis variable across classification levels. **Note:** To create comparative probability plots or comparative Q-Q plots, use the Distribution Analysis task.

P-P Plots: Selecting the Distribution

In the selection pane under the Distributions heading, click **Summary** to access these options.

You can fit one theoretical distribution on a P-P plot. After selecting the distribution, select the distribution name in the selection pane to specify the parameter values and line properties for that distribution.

Option Name	Description
Distributions	Use the check boxes to select one or more of these estimated density curves: ■ **Normal** fits a normal density with a mean (mu) and a standard deviation (sigma). ■ **Lognormal** fits a lognormal density with a scale parameter (zeta), a threshold parameter (theta), and a shape parameter (sigma). ■ **Exponential** fits an exponential density with a scale parameter (sigma) and a threshold parameter (theta). ■ **Weibull** fits a three-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). ■ **Beta** fits a beta density with a scale parameter (sigma), a threshold parameter (theta), and shape parameters (alpha and beta). ■ **Gamma** fits a gamma density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (alpha).
Graphics style	Select the graphics style to use in the output. You can choose between traditional SAS graphics or the ODS statistical graphics. For more information about ODS graphics, see *SAS Output Delivery System: User's Guide*.

P-P Plots: Customizing the Distribution

In the selection pane under the Distributions heading, click the name of the distribution that you want to customize to access these options.

Note: These distribution options are available only if you previously selected the distribution on the **Summary** page or if you select the ***Distribution name*** check box on the page for that distribution.

For the normal, lognormal, exponential, Weibull, beta, and gamma distributions, you can specify the parameters for each analysis variable or all of the analysis variables. By default, the task estimates the parameter values for all of the analysis variables.

To specify the parameters for each analysis variable:

1. Clear the **Apply distribution on all variables** check box.
2. Select the variable's check box in the **Analysis variables** box.
3. Specify the parameter values.

Note: To return to the estimated parameter values, click **Use Estimates**.

Capability Analysis: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Adding an Inset

In the selection pane under the **Appearance** heading, click **Inset** to access these options.

You can enhance a plot by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative histogram, an inset appears in each component histogram.

Select **Include inset** to add summary statistics to the plot. Then select the statistics that you want to include in the inset.

Note: The statistics for calculating the capability indices are available only if you created specification limits for the variable that you assigned to the **Analysis variables** role. You can create specification limits for histograms, CDF plots, probability plots, and Q-Q plots. For more information about calculating capability indices, see the Help for the CAPABILITY procedure.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

The plot appearance options that are available depend on which type of plot the task produces.

Role Name	Description
Background	specifies the background color for the plot.
Bar outline and **Bar fill**	specifies the colors for the bars in a histogram. **Note:** These options are available only for histograms.

Role Name	Description
Reverse the two axes putting the percentiles on the opposite axis	reverses the axes on the plot. If you select this option, then the axes options and reference lines that you selected might no longer work since the orientation of the plot will have changed. **Note:** This option is available only for probability plots and Q-Q plots.
Suppress default legends	prevents the default legend from being displayed with the plot.
Use midpoints	specify midpoints for the histogram. The options that are available from the drop-down list depend on whether you are creating a histogram or a comparative histogram. If you assign a variable to the **Classification variables** role, you are creating a comparative histogram. For comparative histograms, these options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Key** determines the midpoints for the data in the key cell. By default, the levels for the classification variable are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first classification variable. If you specified only one classification variable, then the key cell is the level that occurs first in this order. If you specified two classification variables, then the key cell is the combination of levels of variable 1 and variable 2 that occurs first in this order. Thus, the choice of the key cell determines the uniform horizontal axis that is used for all cells. The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. ■ **Uniform** determines the number of midpoints based on the total sample size. If you do not assign a variable to the **Classification variables** role, you are creating a histogram. For histograms, the following options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Specify number of bins** enables you to specify the number of bins (also called histogram intervals) for the data. By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. **Note:** You can specify midpoints only if you are using the Histograms task.
Specify custom chart size (in pixels)	specifies the image size. For reference, the total screen size is given in the **Width** and **Height** boxes.

Note: Comparative options are available for comparative histograms. You can create a comparative histogram by assigning a variable to the **Classification variables** role in the Histograms task.

Role Name	Description
Rows and **Columns**	If you assign at least one classification variable to an analysis role, then you can use the **Rows per page** and **Columns per page** boxes to change the arrangement of rows and columns in the comparative histogram. If you assign one classification variable, the default arrangement is two rows and one column per page. If you assign two classification variables, the default arrangement is two rows and two columns per page. You can also use the **Frame side** and **Frame top** drop-down lists to specify the colors that are used to fill the frame area of the row labels and the column labels.
Number of bins	By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. You can also specify a standard deviation such that the number of bins that are displayed is based on a range of the standard deviation value above and below the mean of the data.

Setting Specification Limit Options

Note: These options are available for histograms, probability plots, Q-Q plots, and CDF plots.

In the selection pane under the **Appearance** heading, click **Spec Limits** to specify the lower specification limit, the target, and the upper specification limit.

Note: The **Spec Limits** options are available only if you specify one or more limits for at least one analysis variable. To specify a specification limit for an analysis variable, click **Data** in the selection pane, and then select the analysis variable.

Capability Analysis: Specifying the Types of Analyses

In the selection pane, click **Tables** to access these options.

By default, the Capability Analysis report includes the following tables: a basic confidence intervals table, a basic measures table, a tests for location table, and a moments table.

To add a table to the report , select the table in the **Tables** box.

To specify the analysis options, select the table in the **Tables** box. If options are available, a box appears on the right that you use to specify the options.

To suppress all descriptive statistics tables from the report , select the **Suppress descriptive statistics and capability indices tables** check box. Selecting this option does not suppress these tables:

- tables that are created by the distribution that you selected. You can suppress these tables by selecting the **Suppress distribution tables** check box on the panel for that distribution.
- tables that are created by the INTERVALS statement. To suppress these tables, you can specify the NOPRINT option. For more information about the INTERVALS statement, see the Help for the CAPABILITY procedure.

To save your results to an output data set , select the **Save output statistics to a data set** check box. The data set will include univariate statistics and capability indices.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

64

Prepare Time Series Data

About the Prepare Time Series Data Task

The Prepare Time Series Data task can be used to make data more suitable for analysis by other time series tasks. It can also be used to perform generic transformations on data that is intended for use in any other tasks. For example, you can apply mathematical operations, functions, and other transformations to variables in a data set.

Be careful when using this task to manipulate time series data, because the options that you select can affect the accuracy of analyses that are performed on the data later.

You might use this task to perform these tasks:

- interpolate quarterly estimates from an annual series, or aggregate quarterly values to produce an annual series
- combine time series that are measured at different sampling frequencies. For example, suppose that you have data on monthly money stocks, quarterly gross domestic product, and weekly interest rates, and you want to perform an analysis of a model that uses all these variables. To perform the analysis, you first need to convert the series to a common frequency and combine the variables into one data set.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	EXPAND
Required SAS products	Base SAS, SAS/ETS
Recommended additional SAS products	None

Prepare Time Series Data: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Time ID variable	specifies the variable that is used to assign dates to the rows. You must assign exactly one variable to this role. The first variable that contains a SAS date, time, or datetime format is automatically assigned to this role. You can also create a time ID variable.
Time series variables	specifies the variables that you want to prepare for analysis by applying the mathematical operations available in this task. If no variables are specified, then the task converts all numeric variables in the input data, according to the Characteristics options that you set.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for another role.

Prepare Time Series Data: Specifying the Output Frequency

In the selection pane under the Characteristics heading, click **Frequencies** to access these options.

- To leave the output frequency unchanged, select **Same as input frequency**.
- To change the frequency interval, select **Interval** and then select the output interval from the drop-down list.
- To change the ratio of input frequency to output frequency, select **Ratio**. .Use the **Input observations** and **Output observations** boxes to specify the ratio of input observations to output observations.

Prepare Time Series Data: Specifying the Time Series Characteristics

In the selection pane under the Characteristics heading, click **Options** to access these options.

Option Name	Description
Interpolation method	specifies the interpolation method. These methods are available: ■ **Cubic spline** fits a cubic spline curve to the input values. A cubic spline is a segmented function consisting of third-degree (cubic) polynomial functions joined together so that the whole curve and its first and second derivatives are continuous. ■ **Linear spline** fits a continuous curve to the data by connecting successive straight line segments. ■ **Step function** fits a discontinuous piecewise constant curve. ■ **Simple aggregation** performs simple aggregation of time series without interpolation of missing values. ■ **No interpolation** specifies that no interpolation be performed. **Note:** The **No interpolation** option is not available if a frequency conversion has been selected. For more information, see "Prepare Time Series Data: Specifying the Output Frequency" on page 392. For the cubic spline method, you can select the lower and upper endpoint constraints. Here are the endpoint constraint options: ■ **Not-a-knot** means that the first two spline pieces are constrained to be part of the same cubic curve, as are the last two pieces. This is the default. ■ **Natural spline** specifies the natural spline constraint. The second derivative of the spline curve is constrained to be zero at the endpoint. ■ **Slope** specifies the first derivative of the spline curve at the endpoint. ■ **Curvature** specifies the second derivative of the spline curve at the endpoint. For the **Slope** and **Curvature** options, you must specify the first derivative value and the second derivative value, respectively. The default for both derivative values is 1.

Option Name	Description
Observation characteristics	specifies the observation characteristic for the input and output time series. The observation characteristic defines what type of observation each value in the time series is. This ensures that interpolations are performed correctly. Here are the available options: ■ **Total** specifies that the data values represent period totals for the time interval that corresponds to the observation. ■ **Average** specifies that the data values represent period averages. ■ **Beginning** specifies that the data values are beginning-of-period values. This is the default. ■ **Middle** specifies that the data values are period midpoint values. ■ **End** specifies that the data values are end-of-period values. ■ **Derivative** specifies that the output data be the derivatives of the cubic spline curve fit to the input data. Note: The **Derivative** option is available only from the **Output** drop-down list. The difference between the date alignment and the observation characteristics can be confusing. The date alignment option controls the date values that are assigned to the time ID variable for each observation. The observation characteristics control how the time series values are mathematically interpreted when interpolations are performed.
Data alignment in interval	specifies r the alignment of dates within each interval. For example, if you have specified a monthly interval, the **Beginning** option starts the time interval on the first day of the month, the **Middle** option starts the time interval in the middle of the month, and the **End** option starts the time interval on the last day of the month.

Prepare Time Series Data: Specifying Transformations for Each Time Series Variable

In the selection pane under the Transformations heading, click **Before Interpolation** or **After Interpolation** to access these options.

Unique transformations can be applied to each variable that is assigned to the **Time series variable** role.

Option Name	Description
Before interpolation / After interpolation operations	Click **Add** to display a list of available transformations. Select the check box of any transformations that you want to add, and click **OK**. To remove a transformation, select it and then click **Remove**. The transformations are performed in the order in which they appear in the list. You can change the order by selecting a transformation and then using the **Up** and Down buttons to move it in the list. To apply the transformation y = 5 + log(x + 10) to the input time series, you would enter these transformations to the list in this order: 1 Add a constant (n = 10) 2 log 3 3 Add a constant (n = 5)

Option Name	Description
Options for selected operation	Some transformations require that you specify additional options, which will appear in the **Options for selected operation** area. The **Run** and **Save** buttons are not available until all transformation parameters are specified.

Prepare Time Series Data: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Saving output data	The data set that contains the converted or transformed time series variables is always saved. The name of this output data set appears in the **Modified data** box. To permanently save the coefficients of the spline curves that have been fitted to the input time series, select the **Parameter estimates** check box. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Including series plots in the output	Select the **Create plots for input and output series** check box to include plots of the transformed input series, the transformed output series, and the converted series in the output. Plots of the series and the intermediate series are created for each variable that you assigned to the **Time series variable** role.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used

- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

Creating Observation Rows for Missing Rows in the Data Set

Most time series analysis tasks assume that the time series data observations are taken at equally spaced time intervals or periods. If any time intervals are missing, then the data is not equally spaced, and the analysis might fail or produce invalid results. This data is not equally spaced:

Row	Date	Value
1	JAN10	100
2	FEB10	110
3	MAR10	120
4	MAY10	140
5	JUN10	150

This data is not equally spaced, because the interval between observations 3 and 4 is not the same as the interval between observations 2 and 3.

To create observations for the missing intervals:

1. Assign the date variable to the **Time ID variable** role.
2. Set the interval to the frequency of the data. In the example, the data is **Monthly**.
3. Click **Run**.

A new data set is created that contains all the columns in the original data set. New observations for the omitted intervals are added with the values set to missing. This data can now be used in tasks that handle missing values, or the Prepare Time Series Data task can be used on the new data set to interpolate the missing values. For more information, see “Interpolating Missing Values” on page 398.

You can also interpolate the missing values to provide estimates of the missing observations. For more information, see “Converting Aperiodic Series into Periodic Estimates” on page 398.

Converting Time Series Data from One Sampling Frequency to Another

The Prepare Time Series Data task can be used to convert time series data from one sampling frequency to another. For example, weekly data can be aggregated into monthly data, or weekly data can be interpolated from monthly data.

To convert the frequency of your time series data:

1 Assign the date variable to the **Time ID variable** role. In the **Use existing time ID variable** area, set the interval to the frequency of the input data.

2 Assign the time series variables that you want to convert to the **Time series variables** role.

3 In the selection pane under the Characteristics heading, click **Frequencies** and select either **Ratio** or **Interval** in the **Output frequency** area. For example, select **Interval** and set the interval value to **Weekly** to convert the input data to a weekly frequency.

4 In the selection pane under the Characteristics heading, click **Options**, and change the interpolation method to a method that is appropriate for the data. For example, select **Cubic spline** to fit a third-order spline curve to the input data.

5 Select the input data observation characteristic and the desired output data observation characteristic. For example, you can convert beginning-of-month data to quarterly averages.

6 Select the desired output date alignment.

7 Click **Run**.

Changing the Observation Characteristics of the Input Data

The Prepare Time Series Data task can be used to change the observation characteristics of the input data. For example, monthly totals data can be converted to monthly average data.

To change the observation characteristics of your time series data:

1 Assign the date variable to the **Time ID variable** role. In the **Use existing time ID variable** area, set the interval to the frequency of the input data.

2 Assign the time series variables that you want to convert to the **Time series variables** role.

3 In the selection pane under the Characteristics heading, click **Frequencies**. In the **Output frequency** area, leave the output frequency the same as the input frequency unless you are also changing the frequency of the data.

4 In the selection pane under the Characteristics heading, click **Options**, and change the interpolation method to a method that is appropriate for the data. For example, select **Cubic spline** to fit a third-order spline curve to the input data.

5 Select the input data observation characteristic and the desired output data observation characteristic. For example, you can convert beginning-of-month data to quarterly averages.

6 Select the desired output date alignment.

7 Click **Run**.

Converting Aperiodic Series into Periodic Estimates

Most time series analysis tasks assume that the time series data observations are taken at equally spaced time intervals or periods and that no time intervals are missing. Thus, aperiodic series such as transaction data must be converted into periodic data before it is analyzed using the other time series tasks.

To convert aperiodic data into periodic estimates:

1. Assign the date variable to the **Time ID variable** role. In the **Use existing time ID variable** area, select the **Aperiodic data** option.
2. Assign the time series variables that you want to convert to the **Time series variables** role.
3. .In the selection pane under the Characteristics heading, click **Frequencies** and select the desired output interval.
4. In the selection pane under the Characteristics heading, click **Options** and change the interpolation method to a method that is appropriate for the data. For example, select Cubic spline to fit a third-order spline curve to the input data.
5. Select the desired output date alignment.
6. Select the input data observation characteristic and the desired output data observation characteristic. For example, you can convert aperiodic point-in-time values to monthly averages.
7. Click **Run**.

Interpolating Missing Values

The Prepare Time Series Data task can be used to interpolate missing values in the input time series. This data contains missing values:

Row	Date	Value
1	JAN10	100
2	FEB10	110
3	MAR10	120
4	APR10	.
5	MAY10	140
6	JUN10	150
7	JUL10	.
8	AUG10	150

To interpolate the missing values in the input data:

1 Assign the date variable to the **Time ID variable** role. In the **Use existing time ID variable** area, set the interval to the frequency of the input data.

2 Assign the time series variables that you want to convert to the **Time series variables** role.

3 In the selection pane under the Characteristics heading, click **Frequencies** and select **Same as input frequency** for the output series frequency.

4 In the selection pane under the Characteristics heading, click **Options** and change the interpolation method to a method that is appropriate for the data. For example, select **Cubic spline** to fit a third-order spline curve to the input data.

5 Select the input data observation characteristic and the desired output data observation characteristic. For example, you can convert end-of-day data to monthly averages.

6 Select the desired output date alignment.

7 Click **Run**.

Applying Generic Transformations to the Data

The Prepare Time Series Data task can be used to perform a wide variety of mathematical transformations on the input time series. These transformations can be performed in conjunction with the available interpolation operations, or they can be performed without interpolating any values.

To transform data in conjunction with an interpolation operation, follow all of the steps for the interpolation option except for submitting the task. (For more information, see "Interpolating Missing Values" on page 398.) Then use the **Transformations** pages to define the transformations that you want to perform on the data before it is interpolated and after it is interpolated. Either list can be empty.

To transform data without performing any interpolation:

1 Assign the date variable to the **Time ID variable** role. In the **Use existing time ID variable** area, set the interval to the frequency of the input data.

2 Assign the time series variables that you want to convert to the **Time series variables** role.

3 In the selection pane under the Characteristics heading, click **Options** and select **No interpolation** as the interpolation method.

4 Select the desired output date alignment.

5 .In the selection pane under the Transformations heading, click **Before Interpolation** to insert the transformations that you want to perform. Although you would usually leave the **After Interpolation** list empty, you can also define transformations on the **After Interpolation** page. Those transformations will be performed after the transformations in the **Before Interpolation** list are performed.

6 Click **Run**.

65

Principal Components

About the Principal Components Task

Principal component analysis is a multivariate technique for examining relationships among several quantitative variables. You should use principal component analysis if you are interested in summarizing data and detecting linear relationships.

You might use the Principal Components task when you have too many variables to plot simultaneously. For example, suppose that you have the crime rates per 100,000 people in 7 categories for each of the 50 states in 1977. Since there are seven numeric variables, it is impossible to plot all the variables simultaneously. You can use principal component analysis to summarize the data in two or three dimensions and to help you visualize the data.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	PRINCOMP
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Principal Components: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	performs analysis on each variable in this list. You must assign at least one variable to this role.
Partial variables	specifies the variables that you want partialed out when analyzing a partial correlation or covariance matrix. The Principal Components task computes the principal components of the residuals from the prediction of the analysis variables by the variables in this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analyses by a variable that you selected as analysis variables.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the values to use to weight the corresponding values in the analysis variable. You can assign a maximum of one variable to this role.

Principal Components: Setting Analysis Options

In the selection pane, click **Analysis** to access these options.

Option Name	Description
Specifying your analysis	Select the analysis that you want to perform: ■ **Correlations** computes the principal components from the correlation matrix. ■ **Covariances** computes the principal components from the covariance matrix. Specifying this option causes variables with large variances to be more strongly associated with components with large eigenvalues and causes variables with small variances to be more strongly associated with components with small eigenvalues. You should not specify this option unless the units in which the variables are measured are comparable or the variables are standardized in some way. ■ **Uncorrected correlations** computes the principal components from the correlations matrix. However, the intercept is omitted from the model, so the correlations matrix and the standard deviations might not be corrected for the mean. ■ **Uncorrected covariances** computes the principal components from the covariance matrix. However, the intercept is omitted from the model, so the covariance matrix and the standard deviations might not be corrected for the mean.

Option Name	Description
Specifying the number of principal components to be computed	The default setting is the number of variables. Click the up and down arrows to change this number.
Changing the prefix for the principal components	Specify the prefix for the principal components. The name that you specify can have a maximum length of 26 characters. If you do not specify a prefix, then the task uses the default value. By default, the prefix for naming the principal components is PRIN1, PRIN2, ..., PRIN**n**.
Specifying the singularity criterion	Specify the singularity criterion where 0 < p < 1. If a variable that you assigned to the Partial variable role has an R^2 as large as 1-p when predicted from the variables listed before it in the PARTIAL statement, the variable is assigned a standardized coefficient of 0. The default is **1E-08**.
Specifying the divisor for the variance	By default, the divisor for the variance calculation is the degrees of freedom. You can choose an alternative divisor by selecting the method from the **Divisor for variance** drop-down list. **Note:** For partialing, you must assign a variable to the **Partial variable** role.

Principal Components: Generating Plots

In the selection pane, click **Plots** to access these options.

Option Name	Description
Create scree and variance plots	creates a scree plot of eigenvalues and a plot of the proportion variance. You can use a scree plot to decide how many components to use in the analysis. Eigenvalues correspond to each of the principal components and represent a partitioning of the total variation in the sample. The eigenvalues are ordered from largest to smallest, with the first few forming most of the variation. The scree plot generally shows a steep drop over the first few eigenvalues, followed by a leveling off for the remaining ones. You can use the number of eigenvalues in this transition area to determine the appropriate number of components to include.
Create a scatter matrix principal component scores plot	creates a scatter plot matrix of the principal component scores. The histogram of each component is displayed in the diagonal element of the matrix.
Create a pattern profile plot	creates a pattern profile plot. There is one profile for each component. The value on the Y axis is the correlation between the variable and the principal component.
Create a scatter plot of the principal components scores	creates a scatter plot for each principal components score. You can customize these plots by choosing to show the prediction ellipses for the principal component scores of a new observation. By default, no ellipses are shown.
Create pattern component plots	creates pairwise component pattern plots. Each observation on the plot is the correlation between the variable and the two corresponding components on the plot. You can choose whether to plot the patterns in a vector. If you choose to show the vectors, then by default, a unit circle with a 100% variance circle is plotted for the vector pattern plot.

Principal Components: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Saving the output data	saves some or all of your results in an output table. Select the check boxes of the statistics that you want to include in the output. ■ The **Create table containing original data and component scores** option creates a table that contains your input data, plus the principal component scores. ■ The **Create table to contain statistics** option creates a table that contains means, standard deviations, number of observations, correlations or covariances, eigenvalues, and eigenvectors. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Standardizing principal component scores	Select the **Standardize principal component scores** check box if you want to standardize the principal component scores to equal variance. This option affects the table that is specified in the **Create table containing original data and component scores** box.
Suppressing reports	Does not display the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

66

Probability Plots

About the Probability Plots Task

A process capability analysis compares the distribution of output from a process in statistical control to its specification limits in order to determine the consistency with which the specifications are met. A probability plot is used in process capability analysis to compare ordered values of a variable to the percentiles of a specified theoretical distribution such as the normal. If the data distribution matches theoretical distribution, the points on the plot form a linear pattern. Thus, you can use a probability plot to determine how well a theoretical distribution models a set of measurements.

Note: Probability plots are similar to Q-Q plots. Probability plots are preferable for graphical estimation of percentiles, whereas Q-Q plots are preferable for graphical estimation of distribution parameters and capability indices.

You might use a probability plot to analyze a manufacturing process. For example, suppose that the diameters of 50 steel rods are measured and saved in a data set. The process of producing the rods is in statistical control, and you decide to check whether the diameters are normally distributed. The linearity of the point pattern indicates whether the measurements are normally distributed.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CAPABILITY

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Capability Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the numeric variables to be analyzed. You must assign at least one variable to this role. For all plots except the P-P plot, you can specify an upper limit, a target value, and a lower limit in the **Spec limits for *variable-name*** area for each analysis variable. When you specify these limits, the output automatically includes a table that shows the tests for normality, a table that shows the specification limits, and a table that shows the process capability indices.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight. When you assign a variable to this role, the value of the variable in each observation is used to compute weighted statistics. You can assign only one variable to this role. **Note:** Assigning a variable to the **Relative weight** role is not recommended when you are creating a histogram, probability plot, P-P plot, Q-Q plot, or CDF plot. If you assign a variable to this role, any output should be interpreted with caution.
Classification variables (available for the Histogram task only)	specifies one or two variables to use as classification variables. When you assign a variable to this role, the task creates a comparative histogram that enables you to compare the distribution of an analysis variable across classification levels. **Note:** To create comparative probability plots or comparative Q-Q plots, use the Distribution Analysis task.

P-P Plots: Selecting the Distribution

In the selection pane under the Distributions heading, click **Summary** to access these options.

You can fit one theoretical distribution on a probability plot. After selecting the distribution, select the distribution name in the selection pane to specify the parameter values and line properties for that distribution.

Option Name	Description
Distributions	Use the check boxes to select one or more of these estimated density curves: ■ **Normal** fits a normal density with a mean (mu) and a standard deviation (sigma). ■ **Lognormal** fits a lognormal density with a scale parameter (zeta), a threshold parameter (theta), and a shape parameter (sigma). ■ **Exponential** fits an exponential density with a scale parameter (sigma) and a threshold parameter (theta). ■ **Beta** fits a beta density with a scale parameter (sigma), a threshold parameter (theta), and shape parameters (alpha and beta). ■ **Gamma** fits a gamma density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (alpha). ■ **Weibull (3-parameter)** fits a three-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). ■ **Weibull (2-parameter)** fits a two-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). In the two-parameter Weibull, it is assumed that your data has a known lower threshold parameter, theta=0.
Graphics style	Select the graphics style to use in the output. You can choose between traditional SAS graphics or the ODS statistical graphics. For more information about ODS graphics, see *SAS Output Delivery System: User's Guide*.

Probability Plots: Customizing the Distribution

In the selection pane under the Distributions heading, click the name of the distribution that you want to customize to access these options.

Note: These distribution options are available only if you previously selected the distribution on the **Summary** page or if you select the ***Distribution name*** check box on the page for that distribution.

For the normal, lognormal, exponential, Weibull, beta, and gamma distributions, you can specify the parameters for each analysis variable or all of the analysis variables. By default, the task estimates the parameter values for all of the analysis variables.

To specify the parameters for each analysis variable:

1 Clear the **Apply distribution on all variables** check box.

2 Select the variable's check box in the **Analysis variables** box.

3 Specify the parameter values.

Note: To return to the estimated parameter values, click **Use Estimates**.

Capability Analysis: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Adding an Inset

In the selection pane under the **Appearance** heading, click **Inset** to access these options.

You can enhance a plot by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative histogram, an inset appears in each component histogram.

Select **Include inset** to add summary statistics to the plot. Then select the statistics that you want to include in the inset.

Note: The statistics for calculating the capability indices are available only if you created specification limits for the variable that you assigned to the **Analysis variables** role. You can create specification limits for histograms, CDF plots, probability plots, and Q-Q plots. For more information about calculating capability indices, see the Help for the CAPABILITY procedure.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

The plot appearance options that are available depend on which type of plot the task produces.

Role Name	Description
Background	specifies the background color for the plot.

Role Name	Description
Bar outline and **Bar fill**	specifies the colors for the bars in a histogram. **Note:** These options are available only for histograms.
Reverse the two axes putting the percentiles on the opposite axis	reverses the axes on the plot. If you select this option, then the axes options and reference lines that you selected might no longer work since the orientation of the plot will have changed. **Note:** This option is available only for probability plots and Q-Q plots.
Suppress default legends	prevents the default legend from being displayed with the plot.
Use midpoints	specify midpoints for the histogram. The options that are available from the drop-down list depend on whether you are creating a histogram or a comparative histogram. If you assign a variable to the **Classification variables** role, you are creating a comparative histogram. For comparative histograms, these options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Key** determines the midpoints for the data in the key cell. By default, the levels for the classification variable are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first classification variable. If you specified only one classification variable, then the key cell is the level that occurs first in this order. If you specified two classification variables, then the key cell is the combination of levels of variable 1 and variable 2 that occurs first in this order. Thus, the choice of the key cell determines the uniform horizontal axis that is used for all cells. The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. ■ **Uniform** determines the number of midpoints based on the total sample size. If you do not assign a variable to the **Classification variables** role, you are creating a histogram. For histograms, the following options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Specify number of bins** enables you to specify the number of bins (also called histogram intervals) for the data. By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. **Note:** You can specify midpoints only if you are using the Histograms task.
Specify custom chart size (in pixels)	specifies the image size. For reference, the total screen size is given in the **Width** and **Height** boxes.

Note: Comparative options are available for comparative histograms. You can create a comparative histogram by assigning a variable to the **Classification variables** role in the Histograms task.

Role Name	Description
Rows and **Columns**	If you assign at least one classification variable to an analysis role, then you can use the **Rows per page** and **Columns per page** boxes to change the arrangement of rows and columns in the comparative histogram. If you assign one classification variable, the default arrangement is two rows and one column per page. If you assign two classification variables, the default arrangement is two rows and two columns per page. You can also use the **Frame side** and **Frame top** drop-down lists to specify the colors that are used to fill the frame area of the row labels and the column labels.
Number of bins	By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. You can also specify a standard deviation such that the number of bins that are displayed is based on a range of the standard deviation value above and below the mean of the data.

Setting Specification Limit Options

Note: These options are available for histograms, probability plots, Q-Q plots, and CDF plots.

In the selection pane under the **Appearance** heading, click **Spec Limits** to specify the lower specification limit, the target, and the upper specification limit.

Note: The **Spec Limits** options are available only if you specify one or more limits for at least one analysis variable. To specify a specification limit for an analysis variable, click **Data** in the selection pane, and then select the analysis variable.

Capability Analysis: Specifying the Types of Analyses

In the selection pane, click **Tables** to access these options.

By default, the Capability Analysis report includes the following tables: a basic confidence intervals table, a basic measures table, a tests for location table, and a moments table.

To add a table to the report , select the table in the **Tables** box.

To specify the analysis options, select the table in the **Tables** box. If options are available, a box appears on the right that you use to specify the options.

To suppress all descriptive statistics tables from the report , select the **Suppress descriptive statistics and capability indices tables** check box. Selecting this option does not suppress these tables:

- tables that are created by the distribution that you selected. You can suppress these tables by selecting the **Suppress distribution tables** check box on the panel for that distribution.
- tables that are created by the INTERVALS statement. To suppress these tables, you can specify the NOPRINT option. For more information about the INTERVALS statement, see the Help for the CAPABILITY procedure.

To save your results to an output data set , select the **Save output statistics to a data set** check box. The data set will include univariate statistics and capability indices.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

67

Proportional Hazards

About the Proportional Hazards Task

The Proportional Hazards task performs regression analysis of survival data based on the Cox proportional hazards model. This semiparametric model is widely used in the analysis of survival data to explain the effects of explanatory variables on survival times.

Survival analysis is used heavily in clinical and epidemiological follow-up studies. Other fields that use survival analysis methods include sociology, engineering, and economics. Regardless of the field, the common objective of a survival analysis study is to determine not only whether an event occurred, but when it occurred. For example, subjects who die five years after surgery are different from subjects who die one month after surgery. An analysis that simply counts deaths after surgery would ignore valuable information about survival time. Survival analysis can also be used to analyze outcomes other than time. For example, an engineer might want to analyze the amount of mileage until a tire fails or the number of cycles until an engine requires repair. What is common across these studies is that you are analyzing an outcome until an event occurs.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	PHREG
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

Proportional Hazards: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Survival time	specifies the variable to use as the survival time variable. Survival time is required, and you can assign only one variable to this role.
Censoring variable	specifies the variable to use as the censoring variable. Select a value in the **Right Censoring Values** area to create a list of one or more values that are used to censor survival time. Use the **Enter custom values** box to type a specific value and click **Add** to include it in the list of censor values. To remove a value from the list of censor values, clear the check box beside that value. The censoring values should be nonmissing numeric values. The censoring variable is not required. However, you can assign only one variable to this role.
Explanatory variables	specifies the explanatory variables for your model.
Strata variables	specifies the variables whose values determine the stratum levels. If you do not specify endpoints for the numeric variables, then the strata are determined by the combination of levels of the unique values of the strata variables. To allow missing values to form valid stratum levels, click the check box. Otherwise, observations with missing values are deleted from the analysis. For numeric strata variables, you can specify the strata intervals by clicking **Specify intervals in Stratum levels** area. Then define the intervals by adding endpoints. By default, the unique values form the strata levels. If you assign multiple variables to this role, then define the intervals separately for each variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
ID variables	specifies one or more variables to use as ID variables in the output data set that contains the survival estimates. To create the output data set, indicate on the **Results** page that you want to save the survival estimates.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. You cannot group analyses by a variable that you have already selected for an analysis role.

Proportional Hazards: Setting Model Options

In the selection pane, click **Model** to access these options.

Note: These options are available only if you have assigned a variable to the **Explanatory variables** role.

Option Name	Description
Setting the model selection method	Select a method from the Selection method drop-down list to control which explanatory variables appear in the model. If you do not select a method, the full model is used. **Full model (no selection)** This method is the default and provides no model selection capability. All explanatory variables are included in the model. **Forward selection** The forward selection method first estimates parameters for variables that are forced into the model. These variables are the intercepts and the first *n* explanatory variables that appear in the **Variables to include** list. The default for *n* is zero so that no variables are in the model. Next, this method computes the adjusted chi-square statistics for each variable not in the model and examines the largest of these statistics. The *p*-values for these statistics are compared to the significance level that is specified for including a variable in the model. By default, this value is `0.05`. To change this significance level, enter the value in the **To enter the model** box. If no adjusted chi-square statistic has a significance level greater than this value, the forward selection stops. Otherwise, the forward selection method adds the variable that has the largest adjusted chi-square statistic to the model. The forward selection method then calculates adjusted chi-square statistics again for the variables that still remain outside the model, and the evaluation process continues. Thus, variables are added sequentially to the model until no remaining variable produces a significant adjusted chi-square statistic. After a variable is added to the model, it stays there. **Backward elimination** The backward elimination method first estimates parameters for the complete model and eliminates insignificant variables unless explanatory variables are in the **Variables to include** list. Results of the Wald test for individual parameters are examined. The least significant variable that does not meet the significance level that is specified in the **To stay in the model** box is removed. (By default, this value is `0.05`.) After a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified level for removal. **Stepwise selection** The stepwise method is a modification of the forward selection technique except that variables already in the model do not necessarily remain. Variables are added to and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model or if the variable that was just added to the model is the only variable that is removed in the subsequent backward elimination step. **Best subset selection** This selection method orders the models by the likelihood score (chi-square) statistic. It identifies a specified number of models that have the highest-score chi-square statistic for all possible model sizes, ranging from one explanatory variable up to a single model with all the explanatory variables. If there are no more than 10 explanatory variables, then all possible models are listed for each model size. If there are more than 10 explanatory variables, then the number of models that are selected for each model size is, at most, equal to the total number of explanatory variables. The smallest models contain *n* explanatory variables, where *n* ranges from the number of explanatory variables in the **Variables to include** list to the total number of explanatory variables in the model. If the **Variables to include** list is empty, then *n* is `1`. **Note:** Intercept parameters always stay in the model.

Option Name	Description
Specifying the significance levels	For the forward selection method or the stepwise selection method, the significance level that you specify in the **To enter the model** box determines whether an explanatory variable is included in the model. Variables that are not in the model are examined, and the variable that has the smallest *p*-value is entered if the *p*-value is less than or equal to the specified significance level. The default significance level is `0.05`. For the backward elimination method or the stepwise selection method, the significance level that you specify in the **To stay in the model** box determines whether an explanatory variable is kept in the model. Variables in the model are examined, and the variable that has the largest *p*-value is removed if the *p*-value exceeds the specified significance level. The default significance level is `0.05`.
Selecting the variables to stay in the model	The **Variables to include** box lists the explanatory variables that are entered or removed from the model during the variable selection process. To always include a variable in the model, select the check box next to the variable name. If you no longer want to always include a variable in the model, then clear the check box next to the variable name.

Proportional Hazards: Setting the Method Options

In the selection pane, click **Methods** to access these options.

Option Name	Description
Compute confidence limits for hazard ratio	displays the confidence limits for each explanatory variable. You can either enter a value for the confidence level or select a value in the **Confidence level** list. The default value is a 95% confidence level.
Correlations of parameter estimates	displays the estimated correlation matrix of the parameter estimates.
Covariances of parameter estimates	displays the estimated covariance matrix of the parameter estimates.

Option Name	Description
Method for handling failure time ties	specifies how to handle failure time ties. You can choose from these options: ■ **Breslow's approximate likelihood** uses Breslow's approximate likelihood method. If ties are not extensive, then this method provide satisfactory approximations to the computing exact conditional probabilities for the continuous time-scale model. This is the default method. ■ **Discrete logistic model** replaces the proportional hazards model with the discrete logistic model. If the time scale is genuinely discrete, then you should use this method. If the analysis of case-control studies has more than one case in a matched set, then this method is required. ■ **Efron's approximate likelihood** uses Efron's approximate likelihood method. If ties are not extensive, then this method provides satisfactory approximations to computing exact conditional probabilities for the continuous time-scale model. ■ **Compute exact conditional probabilities** computes the exact conditional probability under the proportional hazards assumption that all tied event times occur either before censored times of the same value or before larger values. **Note:** This method can use a considerable amount of computer resources. If there are no ties, all four methods result in the same likelihood and yield identical estimates.

Proportional Hazards: Generating Plots

In the selection pane, click **Plots** to access these options.

Note: Before you can create a model assessment plot, you must assign a variable to the **Explanatory variables** role.

Option Name	Description
Cumulative martingale residual plot	For each explanatory variable, this option plots the observed cumulative martingale residuals against the values of the explanatory variable, along with 20 simulated residual patterns. By selecting the **Panel plot of cumulative martingale residuals** option, you can create a plot with four panels. Each panel contains the observed cumulative martingale residuals and two simulated residual patterns.
Standardized score process plot	For each explanatory variable, this option plots the observed score process component against the follow-up time, along with 20 simulated patterns.
Survival function plot	This option plots the estimated survivor function for a reference set of the explanatory variables. This set consists of the corresponding sample means.
Cumulative hazard function plot	This option plots the estimated cumulative hazard function for a reference set of the explanatory variables. This set consists of the corresponding sample means. If there are no ties, all four methods result in the same likelihood and yield identical estimates.

Proportional Hazards: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Saving the output data	saves some or all of your results in an output table. Select the check boxes of the statistics that you want to include in the output. ■ The **Survival estimates** option creates an output data set that contains survival estimates, diagnostics, and estimated linear predictors. ■ The **Regression coefficient estimates** option creates an output data set that contains the regression coefficient estimates. If you want to include the covariance matrix in the data set, select the **Include estimated covariance matrix** check box. ■ The **Baseline survival function estimates** option creates an output data set that contains the survivor function estimates that correspond to the means of the explanatory variables for each stratum. The data set also includes the estimates of the survivor function and linear predictor, standard errors, confidence intervals, the log of survival, and the log of the negative log of survival. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress all displayed output and plots	Does not display the results. If you select this check box, the options on the **Plots** panel are not available. The results are displayed by default.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

68

Q-Q Plots

About the Q-Q Plots Task

A process capability analysis compares the distribution of output from a process in statistical control to its specification limits in order to determine the consistency with which the specifications are met. A quantile-quantile plot (Q-Q plot) is used in process capability analysis to compare ordered values of a variable to the quantiles of a specified theoretical distribution such as the normal. If the data distribution matches 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.

Note: Q-Q plots are similar to probability plots. Q-Q plots are preferable for graphical estimation of distribution parameters and capability indices, whereas probability plots are preferable for graphical estimation of percentiles.

You might use a Q-Q plot to analyze a manufacturing process. For example, suppose that the distances between two holes that are cut into 50 steel sheets are measured and saved in a data set. The cutting process is in statistical control, and you decide to check whether the process distribution is normal. The Q-Q plot compares the observations in the data set with quantiles of the normal distribution. The linearity of the point pattern indicates whether or not the measurements are normally distributed.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	CAPABILITY

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

Capability Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the numeric variables to be analyzed. You must assign at least one variable to this role. For all plots except the P-P plot, you can specify an upper limit, a target value, and a lower limit in the **Spec limits for *variable-name*** area for each analysis variable. When you specify these limits, the output automatically includes a table that shows the tests for normality, a table that shows the specification limits, and a table that shows the process capability indices.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the variable to use as the relative weight. When you assign a variable to this role, the value of the variable in each observation is used to compute weighted statistics. You can assign only one variable to this role. **Note:** Assigning a variable to the **Relative weight** role is not recommended when you are creating a histogram, probability plot, P-P plot, Q-Q plot, or CDF plot. If you assign a variable to this role, any output should be interpreted with caution.
Classification variables (available for the Histogram task only)	specifies one or two variables to use as classification variables. When you assign a variable to this role, the task creates a comparative histogram that enables you to compare the distribution of an analysis variable across classification levels. **Note:** To create comparative probability plots or comparative Q-Q plots, use the Distribution Analysis task.

Q-Q Plots: Selecting the Distribution

In the selection pane under the Distributions heading, click **Summary** to access these options.

You can fit one theoretical distribution on a Q-Q plot. After selecting the distribution, select the distribution name in the selection pane to specify the parameter values and line properties for that distribution.

Option Name	Description
Distributions	Use the check boxes to select one or more of these estimated density curves: ■ **Normal** fits a normal density with a mean (mu) and a standard deviation (sigma). ■ **Lognormal** fits a lognormal density with a scale parameter (zeta), a threshold parameter (theta), and a shape parameter (sigma). ■ **Exponential** fits an exponential density with a scale parameter (sigma) and a threshold parameter (theta). ■ **Beta** fits a beta density with a scale parameter (sigma), a threshold parameter (theta), and shape parameters (alpha and beta). ■ **Gamma** fits a gamma density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (alpha). ■ **Weibull (3-parameter)** fits a three-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). ■ **Weibull (2-parameter)** fits a two-parameter Weibull density with a scale parameter (sigma), a threshold parameter (theta), and a shape parameter (c). In the two-parameter Weibull, it is assumed that your data has a known lower threshold parameter, theta=0.
Graphics style	Select the graphics style to use in the output. You can choose between traditional SAS graphics or the ODS statistical graphics. For more information about ODS graphics, see *SAS Output Delivery System: User's Guide*.

Q-Q Plots: Customizing the Distribution

In the selection pane under the Distributions heading, click the name of the distribution that you want to customize to access these options.

Note: These distribution options are available only if you previously selected the distribution on the **Summary** page or if you select the ***Distribution name*** check box on the page for that distribution.

For the normal, lognormal, exponential, Weibull, beta, and gamma distributions, you can specify the parameters for each analysis variable or all of the analysis variables. By default, the task estimates the parameter values for all of the analysis variables.

To specify the parameters for each analysis variable:

1. Clear the **Apply distribution on all variables** check box.
2. Select the variable's check box in the **Analysis variables** box.

3 Specify the parameter values.

Note: To return to the estimated parameter values, click **Use Estimates**.

Capability Analysis: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Adding an Inset

In the selection pane under the **Appearance** heading, click **Inset** to access these options.

You can enhance a plot by adding a box or table (referred to as an inset) of summary statistics directly to the graph. If you create a comparative histogram, an inset appears in each component histogram.

Select **Include inset** to add summary statistics to the plot. Then select the statistics that you want to include in the inset.

Note: The statistics for calculating the capability indices are available only if you created specification limits for the variable that you assigned to the **Analysis variables** role. You can create specification limits for histograms, CDF plots, probability plots, and Q-Q plots. For more information about calculating capability indices, see the Help for the CAPABILITY procedure.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

The plot appearance options that are available depend on which type of plot the task produces.

Role Name	Description
Background	specifies the background color for the plot.

Role Name	Description
Bar outline and **Bar fill**	specifies the colors for the bars in a histogram. **Note:** These options are available only for histograms.
Reverse the two axes putting the percentiles on the opposite axis	reverses the axes on the plot. If you select this option, then the axes options and reference lines that you selected might no longer work since the orientation of the plot will have changed. **Note:** This option is available only for probability plots and Q-Q plots.
Suppress default legends	prevents the default legend from being displayed with the plot.
Use midpoints	specify midpoints for the histogram. The options that are available from the drop-down list depend on whether you are creating a histogram or a comparative histogram. If you assign a variable to the **Classification variables** role, you are creating a comparative histogram. For comparative histograms, these options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Key** determines the midpoints for the data in the key cell. By default, the levels for the classification variable are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first classification variable. If you specified only one classification variable, then the key cell is the level that occurs first in this order. If you specified two classification variables, then the key cell is the combination of levels of variable 1 and variable 2 that occurs first in this order. Thus, the choice of the key cell determines the uniform horizontal axis that is used for all cells. The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. ■ **Uniform** determines the number of midpoints based on the total sample size. If you do not assign a variable to the **Classification variables** role, you are creating a histogram. For histograms, the following options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Specify number of bins** enables you to specify the number of bins (also called histogram intervals) for the data. By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. **Note:** You can specify midpoints only if you are using the Histograms task.
Specify custom chart size (in pixels)	specifies the image size. For reference, the total screen size is given in the **Width** and **Height** boxes.

Note: Comparative options are available for comparative histograms. You can create a comparative histogram by assigning a variable to the **Classification variables** role in the Histograms task.

Role Name	Description
Rows and **Columns**	If you assign at least one classification variable to an analysis role, then you can use the **Rows per page** and **Columns per page** boxes to change the arrangement of rows and columns in the comparative histogram. If you assign one classification variable, the default arrangement is two rows and one column per page. If you assign two classification variables, the default arrangement is two rows and two columns per page. You can also use the **Frame side** and **Frame top** drop-down lists to specify the colors that are used to fill the frame area of the row labels and the column labels.
Number of bins	By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display. You can also specify a standard deviation such that the number of bins that are displayed is based on a range of the standard deviation value above and below the mean of the data.

Setting Specification Limit Options

Note: These options are available for histograms, probability plots, Q-Q plots, and CDF plots.

In the selection pane under the **Appearance** heading, click **Spec Limits** to specify the lower specification limit, the target, and the upper specification limit.

Note: The **Spec Limits** options are available only if you specify one or more limits for at least one analysis variable. To specify a specification limit for an analysis variable, click **Data** in the selection pane, and then select the analysis variable.

Capability Analysis: Specifying the Types of Analyses

In the selection pane, click **Tables** to access these options.

By default, the Capability Analysis report includes the following tables: a basic confidence intervals table, a basic measures table, a tests for location table, and a moments table.

To add a table to the report , select the table in the **Tables** box.

To specify the analysis options, select the table in the **Tables** box. If options are available, a box appears on the right that you use to specify the options.

To suppress all descriptive statistics tables from the report , select the **Suppress descriptive statistics and capability indices tables** check box. Selecting this option does not suppress these tables:

- tables that are created by the distribution that you selected. You can suppress these tables by selecting the **Suppress distribution tables** check box on the panel for that distribution.
- tables that are created by the INTERVALS statement. To suppress these tables, you can specify the NOPRINT option. For more information about the INTERVALS statement, see the Help for the CAPABILITY procedure.

To save your results to an output data set , select the **Save output statistics to a data set** check box. The data set will include univariate statistics and capability indices.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

69

Radar Chart

About the Radar Chart Task

The Radar Chart task creates radar (or star) charts that show the relative frequency of data measures in quality control or market research problems. The chart statistics are displayed along spokes that radiate from the center of the chart.

For example, a radar chart can show the total sales for three manufacturing sites. Sales for a site are represented by the length of the spine.

You cannot create a Radar chart if you have selected the Java applet or Java image output formats.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GRADAR
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Radar Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Column to chart	The values of the column that you assign to this role determine the spokes for the radar chart. The column must have at least three observations because it takes three points to define a plane.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value.
Frequency count	The values of the column that you assign to this role provide the frequency (numbers of occurrences) of the values of the **Columns to chart** role. The column must be numeric. If you omit this role, then each value of the **Columns to chart** role is counted exactly once.
Overlay	The values of the column that you assign to this role generate a comparative radar chart by using the levels of the specified column. All charts are displayed in the same set of spokes. You cannot use this role if you assign columns to the **Across** role or **Down** role.
Across	The column that you assign to this role generates a radar chart for each value of the specified column. The charts are drawn in left-to-right order across the graphics area. If you also assign a column to the **Down** role, then the charts are drawn in left-to-right and top-to-bottom order.
Down	The column that you assign to this role generates a radar chart for each value of the specified column. The charts are drawn in top-to-bottom order down the graphics area. If you also assign a column to the **Across** role, then the charts are drawn in left-to-right and top-to-bottom order.

Radar Chart: Setting Appearance Options

Setting Radar Chart Options

In the selection pane under the Appearance heading, click **Radar** to access these options.

Select the style for the chart.

- **Corona** creates a polygon with star vertices that emanate from the inner circle.
- **Polygon** creates a closed polygon.
- **Radial** creates rays that emanate from the center.
- **Spoke** creates rays that emanate from the inner circle.
- **Wedge** creates a closed polygon with rays that emanate from the center to the full spoke length (this is the default).

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run

- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

70

Random Sample

About the Random Sample Task

The Random Sample task creates an output data set that contains a random sample of the rows in the input data set.

You might use this task when you need a subset of the data. For example, suppose you want to audit employee travel expenses in an effort to improve the expense reporting procedure and possibly reduce expenses. You do not have the resources to examine all expense reports, so you want to use statistical sampling to objectively select expense reports for audit.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SURVEYSELECT
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	None

Random Sample: Creating a Random Sample

1 Specify a name for the task.

   Note: This option is available only in SAS Enterprise Guide.

2 Select the input data source for the task.

   By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

3 Specify the variables to include in the output data set. By default, all variables are included in the output data set.

Note: When you assign a variable to the **Strata variable**s role, it is automatically assigned to the **Output variables** role as well. If you remove a variable from the output, it is also removed as a strata variable.

4 Specify the sample size in the desired number of rows or in the desired percentage of input rows. For example, if you specify 3% of rows and there are 400 input rows, then the resulting sample will have 12 rows.

Note: If you assign variables to the **Strata variables** role, this sample size specification applies to each stratus rather than to the entire input data set.

5 Select a sampling method. You can choose from these options:

Simple (no duplicates)
When a row is selected, it is removed from eligibility for subsequent selections. This makes it impossible to select the same row more than once.

Unrestricted (duplicates allowed)
When a row is selected, it remains eligible for subsequent selections. This makes it possible to select the same row more than once. You can specify how multiple selections of the same row are recorded in the output data set. You can choose from the following options:

- **Show each observation once in output (exclude duplicates)** - a row that is selected *n* times will occur in the sample once. In the output, the NumberHits variable (which is calculated automatically by the Random Sample task) lists the number of times that the observation occurred in the input data set.
- **Show all observations in output (include duplicates)** - a row that is selected *n* times will occur in the sample *n* times.

6 Specify the variables to use to partition the input data set into mutually exclusive, nonoverlapping subsets that are known as strata. Each stratus is defined by a set of values of the strata variables, and each stratus is sampled separately. The complete sample is the union of the samples that are taken from all the strata.

Note: If you do not assign any variables to this role, the entire input data set is treated as a single stratus.

You can allocate the total sample size among the strata in proportion to the size of the stratus. For example, the variable GENDER has possible values of M and F, and the variable VOTED has possible values of Y and N. If you assign both GENDER and VOTED to the **Strata variables** role, then the input data set is partitioned into four strata: males who voted, males who did not vote, females who voted, and females who did not vote.

The input data set contains 20,000 rows. Here is the distribution of the data:

- 7,000 males who voted
- 4,000 males who did not vote
- 5,000 females who voted
- 4,000 females who did not vote

Therefore, the proportion of males who voted is 7,000/20,000=0.35 or 35%. The proportions in the sample should reflect the proportions of the strata in the input data set. For example, if your sample data set contains 100 observations, then 35% of the values in the sample must be selected from the males who voted stratus to reflect the proportions in the input data set.

If you select **If there are not enough observations to fill a stratum, then select all of them**, the task selects all the observations in a stratus when the size of the stratus is less than the specified sample size.

7 Specify the name and location for the output data set. By default, the data set is saved to the Work library.

8 Specify the initial seed for the generation of random numbers. If you do not specify a random number seed, then a seed that is based on the system clock will be used to produce the sample.

9 Specify whether to generate a summary table that includes the seed that was used to produce the sample. By specifying this same seed later with the same input data set, you can reproduce the same sample.

71

Rank Data

About the Rank Data Task

The Rank task computes ranks for one or more numeric variables across the observations of a SAS data set and sends the ranks as output to a new SAS data set.

For example, you might want to rank the sales for each product that your company sells. In this case, the ranking variable would show the order of product sales. The product with the highest number of sales would be ranked first.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	RANK
Required SAS products	Base SAS
Recommended additional SAS products	SAS/STAT

Rank Data: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Columns to rank	Each column assigned to this role will be ranked. You must assign at least one variable to this role. By default, the rankings column is given the name rank_*column-name*, where *column-name* is the name of the original column. To specify a new name, enter the name in the **Rank column name** text box.
Rank by	When you assign one or more columns to this role, the input table is sorted by the selected column or columns and rankings are calculated within each group.

Rank Data: Setting Analysis Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Ranking Method	specifies the ranking method. You can choose from these options: ■ **Smallest to largest** ranks the original values from the smallest to the largest value. ■ **Percentile ranks** partitions the original values into 100 groups, in which the smallest values receive a percentile value of 0 and the largest values receive a percentile value of 99. ■ **Deciles** partitions the original values into 10 groups, in which the smallest values receive a decile value of 0 and the largest values receive a decile value of 9. ■ **Quartiles** partitions the original values into 4 groups, in which the smallest values receive a quartile value of 0 and the largest values receive a quartile value of 3. ■ **Group = n (NTILES)** partitions the original values into *n* groups, in which the smallest values receive a value of 0 and the largest values receive a value of *n*-1. Specify the value of *n* in the **Number of groups** text box. ■ **Fractional ranks with denominator = n** computes fractional ranks by dividing each rank by the number of observations that have nonmissing values of the ranking variable. ■ **Fractional ranks with denominator = n+1** computes fractional ranks by dividing each rank by the denominator n+1, where n is the number of observations that have nonmissing values of the ranking variable. ■ **Percents** divides each rank by the number of observations that have nonmissing values of the variable and multiplies the result by 100 to get a percentage. ■ **Normal scores (Blom formula)** computes normal scores from the ranks by using the Blom formula. ■ **Normal scores (Tukey formula)** computes normal scores from the ranks by using the Tukey formula. ■ **Normal scores (van der Waerden formula)** computes normal scores from the ranks by using the van der Waerden formula. ■ **Savage scores (exponential)** computes Savage (or exponential) scores from the ranks.
If values tie, use	specifies the method for resolving ties. If you do not select a method, the task uses the Mean (Midrank) method. ■ **Mean (Midrank)** requests that tied values receive the mean of the corresponding ranks. ■ **High rank** requests that the largest of the corresponding ranks be used. ■ **Low rank** requests that the smallest of the corresponding ranks be used.
Reverse ranking to be from largest to smallest	specifies to reverse the order of the ranking.

Rank: Setting Results Options

In the selection pane, click **Results** to access these options.

By default, the output table contains the original columns as well as the ranked columns. If you want to replace the original column with the ranked columns, then clear the **Include ranking values** check box.

The name of the rankings column is specified when the original column is assigned to the **Columns to rank** role. By default, the rankings column is given the name rank_*column-name*, where *column-name* is the name of the original column. To change the column name, select **Data** in the selection pane.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

72

Recency, Frequency, and Monetary Analysis

About the Recency, Frequency, and Monetary Analysis Task

The Recency, Frequency, and Monetary (RFM) Analysis task is a technique that is used to identify existing customers who are most likely to respond to a new campaign or product offer. This technique is commonly used in direct marketing.

The input data for this task can be transactional data or customer data. The Recency, Frequency, and Monetary Analysis task aggregates the transaction data into customer data. In transactional data, each record represents a transaction.

Here is how the recency, frequency, and monetary scores are determined:

- The recency score is determined by sorting the values of the most recent transaction date in ascending order and then grouping these values into bins. The bin with the oldest dates is assigned a recency score of 1, and so on.
- The frequency score is determined by sorting the values of the number of transactions in ascending order and then grouping these values into bins. The bin with the smallest number of transactions is assigned a frequency score of 1, and so on.
- The monetary score is determined by sorting the values of the total amount of the transaction in ascending order and then grouping these values into bins. The bin with the smallest amount is assigned a monetary score of 1, and so on.

The RFM score is calculated using the following formula: RFM Score = 100 * recency score + 10 * frequency score + monetary score. The least favorable customer segment has the lowest RFM score. The most favorable customer segment has the highest RFM score.

Note: The Recency, Frequency, and Monetary Analysis task requires SAS 9.4 or later.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	
Required SAS products	Base SAS
Recommended additional SAS products	

Recency, Frequency, and Monetary Analysis: Selecting the Type of Data to Analyze

In the selection pane, click **RFM Type** to access these options.

Option Name	Description
Transaction-level data	Each record in the data represents a transaction. The Recency, Frequency, and Monetary Analysis task aggregates the individual transactions by customer ID.
Customer data	Each record represents a customer. Individual transactions are combined before you open the data in the Recency, Frequency, and Monetary Analysis task.

Recency, Frequency, and Monetary Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

The analysis roles that are available depend on the type of data that you want to analyze.

Table 72.1 *Options for Transaction-level Data*

Role Name	Description
Transaction Date	the variable that specifies the date of the transaction. This information determines the most recent transaction date.
Amount of Transaction	the variable that specifies the amount of the transaction. This data is used to compute the total amount of the transactions, which can be the sum, mean, median, or maximum of the amount of the transactions.

Role Name	Description
Customer Identifier	the variable that contains a number or string that can uniquely identify a customer.

Because the Recency, Frequency, and Monetary task aggregates the transactional data into customer data, you must select an aggregation method for the transactions. By default, the aggregation method is Sum.

You can also specify whether to include customers with incomplete transaction records (or records with missing values) in the aggregation.

- If you select this option, the output data set contains the scores for customers with incomplete transaction records. For each customer, only the transaction records that have nonmissing values for the transaction date and the amount of transaction are aggregated. Any records with missing values are ignored.
- If you do not select this option, the output data set does not contain the scores for customers with incomplete transaction records. Transaction records are aggregated only for customers who have complete records.

Table 72.2 *Options for Customer Data*

Role Name	Description
Most Recent Transaction	the variable that specifies the date of each customer's most recent transaction with the company. This data is used in computing the recency score.
Number of Transactions	the variable that specifies the number of transactions for each customer. This data is used in computing the frequency score.
Amount of Transactions	the variable that specifies the total monetary amount spent by each customer. This data is used in computing the monetary score.
Customer Identifier	the variable that contains a number or string that can uniquely identify a customer.

Recency, Frequency, and Monetary Analysis: Categorizing the Data into Bins

In the selection pane, click **Binning** to access these options.

Option Name	Description
Binning methods	specifies the method for categorizing the data. You can choose from these options: ■ The **Nested** method assigns a simple rank to each recency value. Within each recency rank, customers are assigned a frequency rank; and within each frequency rank, customers are assigned a monetary rank. The advantage of this method is that it yields an even distribution of combined RFM scores and takes into consideration that purchase behavior (for example, frequency and monetary) could change over time. The disadvantage of this method is that it takes more time to interpret the frequency and monetary scores, because all the scores are dependent on one another. ■ The **Independent** method independently ranks the recency, frequency, and monetary values. The advantage of this method is that the interpretation of scores is unambiguous. For example, a frequency score of 5 for one customer means the same as a frequency score of 5 for another customer. For smaller samples, the disadvantage of this method is a less even distribution of the combined RFM scores, and the possibility that some segments have no customers.
Number of bins	You can specify the number of bins for recency, frequency, and monetary values. The values that you specify are multiplied to create the total number of bins. For example, if you specify 5 bins for recency, 5 bins for frequency, and 5 for monetary, the total number of bins is 125.

Recency, Frequency, and Monetary Analysis: Saving the Results

In the selection pane, click **Results** to access these options.

Option Name	Description
Location to save RFM scores	By default, recency, frequency, and monetary (RFM) scores are saved in a data set in the Work library. To change this location, click **Edit**.
Output data set variables	You can customize the names of the variables in the output data set.

Recency, Frequency, and Monetary Analysis: Selecting Plots

In the selection pane, click **Plots** to access these options.

Option Name	Description
Monetization map	creates a heat map that summarizes the monetary values for each combination of recency and frequency bin values. This plot is shown by default.

Option Name	Description
RFM distribution charts	creates a bar chart that you can use to check the evenness of the distribution of the bin values. This plot is shown by default.
Summary of RFM segments	creates a table that you can use to check the evenness of the distribution of the bin values.

73

Regression Analysis with Autoregressive Errors

About the Regression Analysis with Autoregressive Errors Task

The Regression Analysis with Autoregressive Errors task estimates and forecasts linear regression models for time series data when the errors are not independent through time (autocorrelated) or the error variance is not constant (heteroscedastic). This task produces forecasts and forecast confidence limits when future values of the independent variables are included in the input data set. It is a useful forecasting tool because it uses the time series part of the model as well as the systematic part in generating predicted values. This model takes into account recent departures from the trend in producing forecasts.

You might use this task to perform these tasks:

- compare estimates from different estimation methods and models in a time series. For example, you can use this task to compare investment models over a period of time.
- create a short-term forecast prediction. Many time series exhibit high positive autocorrelation, because they have the smooth appearance of a random walk. This task can greatly improve the fit of models, not only by adding additional parameters but also by capturing the random walk tendencies.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	AUTOREG
Required SAS products	Base SAS, SAS/ETS

Requirement Name	Procedure and Product Names
Recommended additional SAS products	None

Regression Analysis with Autoregressive Errors: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variable	specifies as the dependent variable for the regression analysis. You must assign exactly one variable to this role.
Explanatory variables	serve as the independent regressor variables for the regression model. If no variables are assigned to this role, only the mean is fitted. (This is a way to obtain autocorrelations of a series.)
Time ID variable	specifies the variable that is used to assign dates to the rows. You must assign exactly one variable to this role. The first variable that contains a SAS date, time, or datetime format is automatically assigned to this role. You can also create a time ID variable. You must assign a variable to this role if you want to generate graphical plots of the analyses.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group by a variable that you have already selected as a dependent variable or as an independent variable.

Regression Analysis with Autoregressive Errors: Setting Model Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Method	specifies the model methods. You can choose from these options: ■ **Maximum likelihood estimates** ■ **Unconditional least-squares estimates** ■ **Yule-Walker estimates** ■ **Iterative Yule-Walker estimates**
Maximum number of iterations	specifies the maximum number of iterations. The default value is `50`.
Use only the first sequence of non-missing data	specifies to use the first contiguous sequence of data with no missing values. If you do not select this option, all complete observations are used.
Include the intercept parameter in the model	specifies to include the intercept parameter in the model. It is included by default.
Fit an autoregressive model	specifies to fit the autoregressive model. You can select or more of these options: ■ Select the **Order of autoregressive process** check box and enter a value in the box or select **Specify lags of autoregressive process** and enter the subset of autoregressive error lags to fit. ■ Select the **Remove nonsignificant autoregressive parameters** check box to remove these parameters from the model in order of least significance (backward elimination). The elimination is done only once on the Yule-Walker estimates that are computed after the initial ordinary least squares estimation. ■ Enter a value in the **Remove if significance level exceeds** box to specify the significance level criterion that is used for backward elimination.

Regression Analysis with Autoregressive Errors: Setting Statistics Options

In the selection pane, click **Statistics** to access these options.

You can select one or more of these statistics.

Option Name	Description
Durbin-Watson statistics	displays Durbin-Watson statistics up to the order *n*. Use the **Include up to order** box to specify n. The range of valid values is 1 to 32,767. The default is `1`.
Marginal probability of Durbin-Watson test	displays *p*-values for the generalized Durbin-Watson test statistics for large sample sizes. Select the **Use linearized approximation of design matrix** option to compute *p*-values that use a linearized approximation of the design matrix when the model is nonlinear due to the presence of an autoregressive error process.

Option Name	Description
Durbin's t statistic	displays the Durbin *t* statistic, which detects residual autocorrelation in the presence of lagged dependent variables. **Note:** If you did not assign a variable to the **Explanatory variables** role, you can calculate only the Durbin *t* statistic.
Durbin's h and t statistics	enables you to choose whether to display the Durbin *h* or *t* statistic. From the **Type** drop-down list, select the statistic to calculate. The Durbin *h* statistic tests for the presence of first-order autocorrelation when the regressors contain the lagged dependent variable. You can select this variable from the **Lagged dependent variable** drop-down list. (The variables in this drop-down list are those variables that you assigned to the **Explanatory variables** role.) If the Durbin *h* statistic cannot be computed, the asymptotically equivalent *t* statistic is displayed instead. **Note:** If you did not assign a variable to the **Explanatory variables** role, you can choose to calculate the Durbin *h* or *t* statistic.
Godfrey's general Lagrange multiplier test against ARMA	displays Godfrey's general Lagrange multiplier test against ARMA up to the order *r*. Use the **Include up to order** box to specify *r*. The range of valid values is 1 to 32,767. The default is **4**.
Jarque-Bera normality test	displays Jarque-Bera's normality test statistic for regression residuals.
Phillips-Perron unit root test	displays the Phillips-Perron unit root test when there are no regressors in the model. To specify truncation points for the weighted variance estimators, select the **Include truncation lags?** check box and enter the lag values in the **Truncation lags** box. Three types of the Phillips-Perron test are reported: zero mean, single mean, and deterministic trend. **Note:** If you assigned a variable to the **Explanatory variables** role, regressors are included in the model, and the Phillips-Ouliaris cointegration test is displayed instead.
Q and LM statistics for absence of ARCH effects	displays a table of statistics that test for heteroscedasticity.
Estimated covariances of the parameter estimates	displays the estimated covariances of the parameter estimates.
Estimated correlations of the parameter estimates	displays the estimated correlations of the parameter estimates.
Partial autocorrelations	displays partial autocorrelations.
Transformation coefficients	displays the transformation coefficients for the first *p* observations. These coefficients are formed from a scalar multiplied by the inverse of the Cholesky root of the Toeplitz matrix of autocovariances.
Inverse of the Toeplitz matrix of the autocovariances	displays the inverse of the Toeplitz matrix of autocovariances for the Yule-Walker solution.
Log likelihood of the regression model	displays the log likelihood value of the regression model, assuming normally distributed errors.

Option Name	Description
Uncentered regression R-square	prints the uncentered regression R^2. The uncentered regression R^2 is useful to compute Lagrange multiplier test statistics, because most LM test statistics are computed as T*URSQ, where *T* is the number of observations that are used in the estimation.
Iteration history of objective function and parameter estim	displays the objective function and parameter estimates at each iteration. The objective function is the full log likelihood function for the maximum likelihood method. The error sum of squares is produced as the objective function of unconditional least squares. For the maximum likelihood estimates method, this option displays the value of the full log likelihood function, not the concentrated likelihood.
Ramsey's RESET Test	displays the Ramsey's RESET test statistics. Note: This option is available only if you assigned a variable to the **Explanatory variables** role.

Regression Analysis with Autoregressive Errors: Generating Plots

In the selection pane, click **Plots** to access these options.

By default, all appropriate plots for the current data selection are included in the output. However, you can choose which plots to include in the output by selecting the **Custom lists of plots** option. You can choose from these options:

Option Name	Description
Autocorrelation function plot	creates a plot of the autocorrelation function.
Inverse autocorrelation function plot of residuals	creates a plot of the inverse autocorrelation function.
Partial autocorrelation function plot of residuals	creates a plot of the partial autocorrelation function.
Predicted and actual values plot	creates a plot of the predicted versus observed values.
Plot Cook's D statistic	creates a plot of Cook's *D* statistic versus the observation number. Note: This option is not available if you selected the **Fit autoregressive model** check box on the Options panel.
Q-Q plot of residuals	creates a normal quantile plot of the residuals.
Plots the residuals	creates a plot of the residual versus predicted values. A residual is the difference between the observed value and the predicted value.

Option Name	Description
Plots the studentized residuals	creates a plot of the studentized residuals versus predicted values. **Note:** This option is not available if you selected the **Fit autoregressive model** check box on the Options panel.
Plots the white noise probabilities	creates a plot of the white noise probabilities.
Plots the histogram of the residuals	creates a histogram of the fit residuals.

Regression Analysis with Autoregressive Errors: Setting Results Options

In the selection pane, click **Resuls** to access these options.

Option Name	Description
Include statistics	saves the observation level statistics. You can choose from these options: ■ **Predicted values, residuals, and confidence limits for the unconditional mean** saves predicted values, residuals, upper confidence limit, and lower confidence limit that are formed from only the structural part of the model. Use the **Confidence levels** drop-down list to set the confidence limit size for the estimates of the structural part of the model. ■ **Predicted values, residuals, and confidence limits for future realizations** saves predicted values, residuals, upper confidence limit, and lower confidence limit that are formed from both the structural and autoregressive parts of the model. Use the **Confidence level**s drop-down list to set the confidence limit size for the estimates of future values of the response time series. ■ **Cumulative sum (CUSUM) and confidence levels** saves the CUSUM statistic, CUSUMSQ statistics, and their upper and lower confidence limits. Use the **Confidence levels** drop-down list to set the significance level for the upper and lower limits. ■ **Recursive and BLUS residuals** saves the values of Theil's BLUS residuals, the recursive residuals that are used to compute the CUSUM and CUSUMSQ statistics, and the part of the predictive error variance (vt) that is used to compute the recursive residuals.
Include parameter estimates	create a data set that contains all the variables that are used in the model. Each regressor variable contains the estimate for the corresponding regression parameter. To include the covariance matrix for the parameter estimates, select the **Include covariance matrix** check box.

Option Name	Description
Suppress all displayed output	does not display the results. **Note:** SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

74

Regression Analysis of Panel Data

About the Regression Analysis of Panel Data Task

The Regression Analysis of Panel Data task analyzes a class of linear econometric models that commonly arise when time series and cross-sectional data are combined. This task deals with panel data sets that consist of time series observations on each of several cross-sectional units.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	TSCSREG
Required SAS products	Base SAS, SAS/ETS
Recommended additional SAS products	None

Regression Analysis of Panel Data: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Dependent variable	specifies the variable to be regressed. You must assign exactly one variable to this role.
Explanatory variable	specifies the independent variables to regress the response variable. You must specify at least one variable to this role.
Cross-section ID variable	specifies the variable that defines the cross-sections. You must assign exactly one variable to this role. The data are automatically sorted by the values of this variable.
Time ID variable	specifies the variable that is used to assign time periods to the observations in each cross section. You must assign exactly one variable to this role. The data are automatically sorted by the values of this variable.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables.

Regression Analysis of Panel Data: Setting Model Options

In the selection pane, these options appear under the **Options** heading.

Plot Name	Description
Estimation methods to use	To specify estimation methods, select the option buttons for the desired methods. The performance of any estimation method for the model regression parameters depends on the statistical characteristics of the error components in the model. The default is **Two-way random effects**. ■ A **One-way fixed effects** model is the specification that depends only on the cross section to which the observation belongs. ■ A **Two-way fixed effects** model is the specification that depends on both the cross section and the time series to which the observation belongs. ■ A **One-way random effects** model is the specification that depends only on the cross section to which the observation belongs. In this case, the effects are random. For a balanced panel, the Fuller and Battese method is used by default. For an unbalanced panel, the Wansbeek and Kapteyn's method is used by default. ■ A **Two-way random effects (variance components model)** model is the specification that depends on the cross section and the time period to which the observation belongs. In this case, the effects are random. For a balanced panel, the Fuller and Battese method is used by default. For an unbalanced panel, the Wansbeek and Kapteyn's method is used by default. ■ An **Autoregressive model (Parks method)** assumes a first order autoregressive error structure with contemporaneous correlation between cross-sections. If you select Parks method, then you can use a cubic spline curve to interpolate the missing values in the series. A cubic spline is a segmented function consisting of third-degree (cubic) polynomial functions joined together so that the whole curve and its first and second derivatives are continuous. A cubic spline curve is fit to the nonmissing values of the variables. This curve is used to form the continuous time approximations for the input series, and the output time series are generated from the spline approximations. By default, the **Cubic spline interpolation of missing values** check box is selected. **Note:** Use the **Cubic spline interpolation of missing values** option with care. Interpolation often introduces bias into the model, which can result in poor parameter estimates and hypothesis tests. For more information about these methods, see *SAS/ETS: User's Guide*.
Include the intercept parameter in the model	To exclude the intercept parameter from the model, clear the check box for this option. By default, the check box is selected.

Regression Analysis of Panel Data: Setting Results Options

In the selection pane, click **Results** to access these options.

To save the parameter estimates to an output table, select the **Save parameter estimates** check box. If desired, select the check boxes to include the **Covariance matrix**, **Correlation matrix**, or both.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

To display the results, select the **Show analysis** check box. To suppress the results, clear this check box. The results are displayed by default.

If you choose to display your results, then select the check boxes for the items that you want to appear. Some of the items are available only if you have chosen the Parks estimation method.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

75

SAS ODS Statistical Graphs

About the SAS ODS Statistical Graphs

SAS ODS Statistical Graphics, more commonly referred to as SAS ODS Graphics, is an extension of the SAS Output Delivery System (ODS). ODS manages all output that is created by procedures and enables you to display the output in a variety of forms, including HTML and PDF.

Many SAS analytical procedures use ODS Graphics functionality to produce graphs as automatically as these procedures produce tables. ODS Graphics use the Graph Template Language (GTL) syntax, which provides the power and flexibility to create many complex graphs. The GTL is a comprehensive language for defining statistical graphics.

In SAS Enterprise Guide, you can use the ODS Graphics Designer to define these statistical graphics without knowing the GTL. After a graph definition is created, you can use that graph definition to create an ODS statistical graph in SAS Enterprise Guideor the SAS Add-In for Microsoft Office.

Using the ODS Graphics Designer

Note: The SAS ODS Graphics Designer is not available in the SAS Add-In for Microsoft Office, but you can open SGD files that were created using SAS ODS Graphics Designer in the SAS add-in.

The SAS ODS Graphics Designer is an interactive graphical application that you can use to create and design custom graphs. The designer creates graphs that are based on the Graph Template Language (GTL), which is the same language that is used by SAS analytical procedures and SAS ODS Graphics procedures. The ODS Graphics Designer provides a graphical user interface, so you can design graphs easily without knowing the details of templates and the GTL.

Note: Beginning with the first maintenance release of SAS 9.4, you can use the SAS ODS Graphics Designer in SAS Enterprise Guide without having Base SAS installed on your local machine.

Using point-and-click interaction, you can create simple or complex graphical views of data for analysis. The ODS Graphics Designer enables you to design sophisticated graphs by using a wide array of plot types. You can design multi-cell graphs, classification panels, and scatter plot matrices. Your graphs can have titles, footnotes, legends, and other graphics elements. You can save the results as an image for inclusion in a report or as an ODS Graphics Designer file (SGD) that you can later edit. For more information, see “Show ODS Statistical Graph” on page 460.

To open the ODS Graphics Designer, select **Tasks** ⇨ **Graphs** ⇨ **Open ODS Graphics Designer**. For more information, see *SAS Output Delivery System: User's Guide*.

Show ODS Statistical Graph

The Show ODS Statistical Graph task is available in both SAS Enterprise Guide and the SAS Add-In for Microsoft Office. In this task, you select the graph definition (SGD file) to use for a selected data source. You can create a graph definition in SAS Enterprise Guide. For more information, see "Using the ODS Graphics Designer" on page 459.

After a graph definition is created, it can be used in local and remote SAS sessions. You can even render these graphs in SAS stored processes.

To use a graph definition (SGD file) to create an ODS statistical graph:

1. Select **Tasks** ⇨ **Graphs** ⇨ **Show ODS Statistical Graph**. The Show ODS Statistical Graph dialog box appears.
2. Select the SGD file that contains the design that you want to use.
3. Select the data source for the graph.
4. (Optional) Specify the width and height for the chart.
5. (Optional) Specify a title and footnote.
6. Click **Run**.

For more information about graph definitions, see *SAS Output Delivery System: User's Guide*.

76

SAS Rapid Predictive Modeler

About the SAS Rapid Predictive Modeler

SAS Rapid Predictive Modeler is designed to build models for the following types of data mining classification and regression problems:

- classification models that predict the value of a discrete variable. Some examples are classification models that predict the value of a variable, such as True or False; Purchase or Decline; High, Medium, or Low; and Churn or Continues.
- regression models that predict the value of a real number variable. Some examples are regression models that might predict amounts such as revenue, sales, or success rate by using real values.

To create a model by using the SAS Rapid Predictive Modeler, you must supply a data set, where every row contains a set of independent predictor variables (known as inputs) and at least one dependent variable (known as a target). The SAS Rapid Predictive Modeler decides whether variables are continuous or categorical, and chooses the input variables that should be included in the model. For more information about how the task samples the data, see Sampling Strategies for the SAS Rapid Predictive Modeler.

Your model can be saved as SAS code and then deployed in a SAS environment. You can use the SAS model code to score new data, and then use the results to make more informed business decisions. This process is called model scoring. You can use scored data to decide which customers to select for certain offers, to anticipate manufacturing, warehousing, or logistics demands, to manage customer churn, or to detect transactions that might be fraudulent.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	
Required SAS products	SAS Enterprise Miner
Recommended additional SAS products	

Sampling Strategies for the SAS Predictive Modeler

The SAS Rapid Predictive Modeler uses a composite sampling approach. The number of observations that are included in the data sample depend on these factors:

- number of input variables
- total number of observations in the data source
- whether the data contains rare event targets
- number of events in the data

Here are the guidelines that the SAS Rapid Predictive Model task uses to determine the number of observations that are processed:

Number of Input Variables	Number of Observations Processed
< 100	80,000
100 – 200	40,000
> 200	20,000

Condition	Rare Event: Yes	Rare Event: No
total number of observations < number of observations being processed OR total number of events < (0.10*number of observations being processed	Sample the data so that there is a 10:1 ratio of non-events to events	no sampling
total number of events > (0.10*number of observations being processed)	Sample the following proportion of the rare events: 10*[(0.10*(number of observations being processed))/number of events]	stratified sampling

SAS Rapid Predictive Modeler: Organizing Data

Before you can build a model, you need input data that represents historical events and characteristics that can be used for prediction. You also need target data that represents the event or value that you want to predict. In many cases, the input data is derived from one time period and the target data is derived from a later time period. The combined input and target data that you use to develop your model is called training data.

For example, you might mine last year's sales receipts to predict next year's expected revenue or to predict which customers will respond to a special offer. Using historical data from past events to predict performance on future events is called model training.

For best model results, your model training data should contain a large number of observations stored as rows of data. For example, many retail customer models use input data that has tens of thousands of observations.

If your target variable contains a rare event (for example, an offer that perhaps only 1% of your customers will respond to), you must ensure that your training data contains a significant number of these customers in your data set. You might want to oversample your training data to make sure you select all customers who accepted the offer, and provide an equal number of customers who did not accept. Oversampling makes it easier for a model with a rare event target to find a stable solution.

When you perform oversampling to boost rare event occurrences in your training data, you artificially inflate the occurrence of targeted events in your training data relative to the natural population. To compensate for the difference between the training data and the population data, the SAS Rapid Predictive Modeler provides you with a prior probability setting. Prior probability settings specify the true proportional frequencies of the targeted event in the population data.

The data that you mine using the SAS Rapid Predictive Modeler should be organized into rows (observations) and columns (variables). One of the columns should represent a target variable. Consider the following example:

Name	Age	Gender	Income	Treatment	Purchase
Ricardo	29	M	33000	Y	Y
Susan	35	F	51000	Y	N
Jeremy	49	M	110000	N	Y

Name
: a column that contains ID values for each observation. The SAS Rapid Predictive Modeler does not process ID variable columns for analytical content.

Age, Gender, Income, and Treatment
: input columns that are used by the SAS Rapid Predictive Modeler.

Purchase
: a target column.

When you configure your table of input data, you can also designate a frequency column. Frequency variables can represent the relative weight that should be assigned to a row (or observation). For example, in some data sets, a single row might represent more than one observation. You can also select columns that you want the SAS Rapid Predictive Modeler to ignore during your analysis.

Training data always requires input and target variable values. Data that you use for scoring requires only input variable values; a target column is optional. When the model is used to make predictions from new data, the target column is not required. When the model is used to monitor effectiveness, the target column is required. Data that you use for scoring also typically includes an ID column.

SAS Rapid Predictive Modeler: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

The analysis roles that are available depend on the type of data that you want to analyze.

Role Name	Description
Dependent variable	specifies the value that you want to predict or classify. The dependent variable is also known as the target variable. You must assign at least one variable to this role.
Frequency count	specifies the variable to use to represent the weightings that you want to apply to rows of data. You can assign only one variable to this role.
ID	specifies variables that are useful for reporting and scoring selection functions. These variables are not included in the analytic analysis.
Excluded	specifies the variables that you do not want to include in your analytic analysis.

SAS Rapid Predictive Modeler: Choosing a Model

Setting the Model Options

In the selection pane, click **Model** to access these options.

With these options, you can specify the complexity level of the model that you want to build. The modeling methods are in a hierarchy: the intermediate method includes basic and intermediate models, and the advanced method includes basic, intermediate, and advanced models.

The models that you create using the basic method will probably run faster than the models that you create using the intermediate method, but the basic method also might create a less accurate model. The same is often true when you compare the models that you create with the intermediate and advanced methods. For more information about the data mining operations that occur for each model, “Understanding the Models for the SAS Rapid Predictive Modeler” on page 466.

SAS Enterprise Miner modeling functions are executed when you run the SAS Rapid Predictive Modeler. The modeling functions that the software runs depend on the selected modeling method.

For the dependent variable, click **Decision and Priors** to specify this information:

Table 76.1 Decision and Priors Options for Dependent Variable

Option Name	Description
Event Level	specifies the class target value that you want to model. The SAS Rapid Predictive Modeler builds a model that provides the probabilities for each target event, but reporting improves when the desired target level is known.
Prior Probabilities	displays the counts and proportions of the target variable levels that occur in the model training data. You can adjust these values when your target variable is a categorical variable, and the training data and population data have different target distributions. For example, consider a model that was trained on oversampled data where 50% of observations are responders and 50% of observations are non-responders. However, the population data that the model scores historically contains only 10% responders and 90% non-responders. You can use prior probability settings to inform the model of the historically expected proportions of responders to non-responders. ■ If you do not want to specify prior probabilities, select **None** (which is the default). ■ To specify equal probabilities for all levels of the target variable, select **Equal**. ■ To specify your own custom prior probabilities for target variable levels in the scored data, select **User Defined** and specify the probabilities. The prior probabilities that you specify must sum to 1.
Decision function	specifies the costs, profits, or weights that you want to associate with the predicted results. The table of values is called a decision matrix. You use a decision matrix to associate a value with each possible decision outcome. ■ If your model does not require a decision matrix, select **None** (which is the default). ■ To use your model to maximize profit, select **Maximum**, and if desired, enter a higher weight in the true positive cell of the matrix. ■ To use your model to minimize cost, select **Minimum**, and if desired, enter a higher weight in the true negative cell of the matrix. ■ To use your model to predict rare events, select **Inverse** to identify true positive and true negative predictions, at the risk of misestimating false positive and false negative predictions.

Table 76.2 Modeling Options

Option Name	Description
Basic	specifies the value that you want to predict or classify. The dependent variable is also known as the target variable. You must assign at least one variable to this role.
Intermediate	specifies the variable to use to represent the weightings that you want to apply to rows of data. You can assign only one variable to this role.
Advanced	specifies variables that are useful for reporting and scoring selection functions. These variables are not included in the analytic analysis.

Understanding the Models for the SAS Rapid Predictive Modeler

The SAS Rapid Predictive Modeler provides you with basic, intermediate, and advanced models. The models increase in sophistication and complexity.

- The basic model is a simple regression analysis.
- The intermediate model includes a more sophisticated analysis, plus the analysis from the basic model, and chooses the better model.
- The advanced model includes an even more sophisticated analysis, plus the analyses from the basic and intermediate model, and chooses the best model.

Basic Model

The basic model performs a series of three data mining operations.

- Variable Selection: The basic model chooses the top 100 variables for modeling.
- Transformation: The basic model performs an Optimal Binning transformation on the top 100 variables selected for modeling. The Optimal Binning transformation compensates for missing variable values, so missing value imputation is not performed.
- Modeling: The basic model uses a forward regression model. The forward regression model chooses variables one at a time in a stepwise process. The stepwise process adds one variable at a time to the linear equation until the variable contributions are insignificant. The forward regression model seeks to exclude variables with no predictive ability (or variables that are highly correlated with other predictor variables) from the analytic analysis.

Intermediate Model

The intermediate model performs a series of seven data mining operations.

- Variable Selection: The intermediate model chooses the top 200 variables for modeling.
- Transformation: The intermediate model performs a best power transformation on the 200 variables that were selected for modeling. The best power transformations are a subset of the general class of transformations that are known as Box-Cox transformations. The best power transformation evaluates a subset of exponential power transformations, and chooses the transformation that has the best results for the specified criterion.
- Imputation: The intermediate model performs an imputation to replace missing variables with the average variable values. The imputation operation also creates indicator variables that enable observations that contain imputed variable values to be identified.
- Variable Selection: The intermediate model uses the chi-square and R-square criteria tests to remove variables that are not related to the target variable.
- Union of Variable Selection Techniques: The intermediate model merges the set of variables that were selected by the chi-square and R-square criteria tests.
- Modeling: The intermediate model submits the training data to three competing model algorithms. The models are a decision tree, a logistic regression, and a stepwise regression. In the case of the logistic regression model, the training data is first submitted to a decision tree that creates a NODE_ID variable that is passed as input to the regression model. The NODE_ID variable is created to enable variable interaction models.
- Champion Model Selection: The intermediate model performs an analytic assessment of the predictive or classification performance of the competing models. The model that demonstrates the best predictive or

classification performance is selected to perform the modeling analysis. The intermediate model for champion model selection evaluates the performance of not only the intermediate models, but also the basic models.

After the SAS Rapid Predictive Modeler chooses the intermediate champion model, it compares the predictive performance of the intermediate champion model to the basic model, and then chooses the best model as the result.

Advanced Model

The advanced model performs a series of seven data mining operations.

- Variable Selection: The advanced model chooses the top 400 variables for modeling.
- Transformation: The advanced model performs the multiple transformation algorithm on the 400 variables that were selected for modeling. The multiple transformation operation creates several variable transformations that are intended for use in later variable selections. Multiple transformations result in an increase in the number of input variables. Because of the increase in input variables, SAS Rapid Predictive Modeler selects the best 400 input variables from the output that was generated by the multiple transformation algorithm.
- Imputation: The advanced model performs an imputation to replace missing variables with the average variable values. The imputation operation also creates indicator variables that enable the user to identify observations that contain imputed variable values.
- Variable Selection: The advanced model uses the chi-square and R-square criteria tests to remove variables that are not related to the target variable. AOV16 variables are created during the R-square analysis.
- Union of Variable Selection Techniques: The advanced model merges the set of variables that were selected by the chi-square and R-square criteria tests.
- Modeling: The advanced model submits the training data to four competing model algorithms. The models are a decision tree model, a neural network model, a backward regression model, and an ensemble model. The neural network model conducts limited searches in order to find an optimal feed-forward network. Backward regression is a linear regression model that eliminates variables by removing one variable at a time until the R-squared scores drop significantly. The ensemble model creates new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor input models. The new ensemble model is then used to score new data. The ensemble model that you use in the advanced model is created from the output of the basic model, the champion model from the intermediate model, and the champion model from the advanced model.
- Champion Model Selection: The advanced model performs an analytic assessment of the predictive or classification performance of the competing decision tree, neural, and regression models. The model that demonstrates the best predictive or classification performance is then used as an input, along with the champion model from the basic and intermediate models, to create an ensemble model. Then the newly created advanced ensemble model, decision tree model, neural model, and backward regression model are analytically compared to select the best model from the sample space of all basic, intermediate, and advanced champion models.

After the SAS Rapid Predictive Modeler selects a champion model, it runs and compares the predictive performance of the advanced model to the champion models for the intermediate and basic models, and then chooses the best performing champion model as the result.

SAS Rapid Predictive Modeler: Setting the Report Options

Report Options

In the selection pane, click **Report** to access these options.

The reports identify significant terms in the model and generate common business graphics, such as lift charts. The results include statistics for training and validation data. The SAS Rapid Predictive Modeler process divides the input data into training data and validation data. Training data is used to compute the parameters for each model, resulting in the training fit statistics. Validation data is then scored with each model, resulting in the validation fit statistics. The validation fit statistics are used to compare models and detect overfitting. If the training statistics are significantly better than the validation statistics, then you would suspect overfitting, which occurs when the model is trained to detect random signals in the data. Models with the best validation statistics are generally preferred.

The SAS Rapid Predictive Modeler automatically generates a concise set of core reports that provide a summary of the data source and variables that were used for modeling, a ranking of the important predictor variables, multiple fit statistics that evaluate the accuracy of the model, and a model scorecard. For more information, see About the standard reports for the SAS Rapid Predictive Modeler.

You can also choose to include this information in your results:

- **Model Summarization**
- **Variable Ranking**
- **Cross Tabulations**
- **Classification Matrix**
- **Fit Statistics**
- **Lift Plot**
- **Model Comparison**

  Note: This option is not available for the Basic method, which uses only one model algorithm.

About the Standard Reports for the SAS Rapid Predictive Modeler

Here are the standard reports that are automatically generated by the SAS Rapid Predictive Modeler:

Gains Chart
: Gains chart plots are available only for models that have class target variables. This chart shows percentiles of the data ranked by predicted value. Lift is a measure of the ratio of the number of target events that the model identified, compared to the number of target events that were found by random selection.

Receiver Operating Characteristic Plot (ROC)
: The Receiver Operating Characteristic plot shows the maximum predictive power of a model for the entire sample (rather than for a single decile). The data is plotted as sensitivity versus (1 – specificity). The separation between the model curve and the diagonal line (which represents a random selection model) is called the Kolmogorov-Smirnov (KS) value. Higher KS values represent more powerful models.

Scorecard
: The results include a scorecard so that the model's characteristics can be interpreted for business purposes. When the software builds a scorecard, each interval variable is binned into distinct ranges of values. Then, each variable is ranked by model importance and scaled to a maximum of 1,000 points. The distinct value for each variable then receives a portion of the scaled point total.

Project Information
: The project information shows which user created the model, when the model was created, and where the model's component files are stored.

SAS Rapid Predictive Modeler: Setting the Output Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Save Enterprise Miner project data	specifies whether to save the SAS Enterprise Miner data from this task. A model from the SAS Rapid Predictive Modeler is an example of a SAS Enterprise Miner project. When you save SAS Enterprise Miner data, you can use the SAS Enterprise Miner interface to open and edit the model that you created using the SAS Rapid Predictive Modeler. In SAS Enterprise Miner , you can save and export your analysis for use outside of SAS Enterprise Miner , and register your model with a SAS Metadata Repository. The **Save Enterprise Miner Project Data** check box is selected by default. ■ If you have a workstation installation of SAS Enterprise Miner, the project is saved locally in the workstation location that is specified in the **SAS App Server Folder** field. Click **Browse** to change this location. ■ If you have a client/server installation of SAS Enterprise Miner, the project is saved in the User Root folder (for example, `C:\users\username\Documents`) on the server. If you used SAS Management Console to define a SYSTEM root location (for example, `C:\projects`) on the server, the **SAS App Server Folder** field is blank. In this case, click **Browse** to specify the location for your project storage. If the **SAS App Server Folder** field is left blank and you do not specify a location before attempting to save the project, an error message is displayed that prompts you to specify a valid location.
Score a data set	specifies the name of the data set that you want to be scored by the model that the SAS Rapid Predictive Modeler builds.

SAS Rapid Predictive Modeler: Registering Your Model

In the selection pane, click **Registeration** to access these options.

About Registering a Model

Models that are created using the SAS Rapid Predictive Modeler are easy to deploy. They are saved as SAS code that you can run on any SAS 9.2 (or later) installation.

Note: You must run or save the SAS Rapid Predictive Modeler at least once before you can register your model with the SAS Metadata Repository. When you created the model, you must select the **Save Enterprise Miner Project Data** option for the registration options to be available.

When you register a model, other SAS users at your site can use the registered model.

- Users of **SAS Enterprise Guide** and the **SAS Add-In for Microsoft Office** can use the model to perform their own scoring processes. For example, a registered model can be used to score new data with the Model Scoring task.
- Users of SAS Enterprise Miner can open these models as projects and diagrams in SAS Enterprise Miner, and modify the models to perform further diagnostics, or investigate behaviors and alternatives.
- Users of **SAS Model Manager** can import the model to be managed with other modeling assets, as well as to monitor model performance for degradation over time.
- Users of SAS Data Integration Studio can import the model to create managed and scheduled scoring processes.

Models that are created using the SAS Rapid Predictive Modeler can also be published using the SAS Scoring Accelerator for Teradata, DB2, or Netezza.

Register a Model to the SAS Metadata Repository

1. Specify the model name, path, and description.
2. Click **Register**.

   Note: The **Register** button is not available if you are not connected to a SAS Workspace Server.

77

Scatter Plot

About the Scatter Plot Task

The Scatter Plot task creates two-dimensional scatter plots, three-dimensional scatter plots, or three-dimensional needle plots that show the relationships between two or three variables by revealing patterns or concentrations of data points.

For example, a two-dimensional scatter plot can display the weights and ages of all patients who are included in a clinical study.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GPLOT, G3D
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Scatter Plot: Selecting a Plot Type

In the selection pane, click **Scatter Plot** to access these options.

Role Name	Description
2D Scatter Plot	creates a two-dimensional plot that displays the values of two variables as data points on one horizontal axis (X) and one vertical axis (Y). Each pair of X and Y values forms a data point.
3D Scatter Plot	creates a three-dimensional plot that displays the values of three numeric variables as data points. Each group of X, Y, and Z values forms a data point.
3D Scatter Plot with Needles	creates a three-dimensional plot that displays the values of three numeric variables as data points. Each group of X, Y, and Z values forms a data point. Each data point is connected to the horizontal plane with a line called a needle.

Scatter Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart. You can create a statistical summary of the unique values of the horizontal variable. Select the variable under the **Vertical** role. Select **Summarize for each distinct horizontal value** and select the summary function. For example, your input data lists the sales and expenses for each country by region. You assign the Country variable to the **Horizontal** role, and the Sales variable to the **Vertical** role. If you want to plot the sum of the sales for each country, select the check box, and then select **Sum** from the drop-down list. The task adds the regional sales for a country and presents a single sales value for each country in the final output.

Role Name	Description
Vertical (Right)	The column that you assign to this role is the vertical or Y axis variable for the right side of the graph. This variable is plotted against the X axis variable so that another plot is produced on the graph. You can create a statistical summary of the unique values of the horizontal variable. Select the variable under the **Vertical** role. Select **Summarize for each distinct horizontal value** and select the summary function. For example, your input data lists the sales and expenses for each country by region. You assign the Country variable to the **Horizontal** role, and the Expenses variable to the **Vertical (Right)** role. If you want to plot the sum of the expenses for each country, select the check box, and then select **Sum** from the drop-down list. The task adds the regional expenses for a country and presents a single expense value for each country in the final output. **Note:** This role is available only for a two-dimensional scatter plot.
Depth	The column that you assign to this role is the third-dimension or Z-axis variable for the chart. **Note:** This option is available only for three-dimensional plots.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. Select the **Produce same axis scaling for all graphs** check box if you want the scale for each axis to be the same for all graphs that are produced. By default, the range of values for each axis is based on the minimum and maximum values in the data that is being plotted, and therefore it could vary from graph to graph. If you select this option, then the range for each axis is the same for all the graphs.

Scatter Plot: Setting Appearance Options

Setting Three-dimensional Scatter Plot Options

In the selection pane under the Appearance heading, click **3D Scatter** to access these options.

Select a symbol type, color, and size, and whether to show needles for each data point.

Use the sliders or the **Rotate** and **Tilt** controls to set the viewing angle. By default, the rotate angle is 70, and the tilt angle is 70. Click **Reset** to return the viewing angle to the default setting.

Note: If you clear the **Specify rotate and tilt values** check box, the rotate and tilt of the graph is determined by the graph output format. By default, this option is selected, and you can specify the viewing angle.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Plots** to access these options.

Note: These options are available only for a two-dimensional scatter plot.

To control the appearance of the plot for a specific variable:

1. Select a variable from the list. The variables in the list are the variables that you assigned to the **Vertical** and **Vertical (Right)** roles. I
2. In the **Line** area, specify the line style, the line width, and the color of the line.

3 Select an outline color. This option specifies the color for these parts:

- the outline of any enclosed areas
- confidence limits in a regression analysis
- the lines or outlines for standard deviation plots

4 In the **Data point marker** area, specify the type of marker, the symbol to use, the height of each marker, and the color for the data point markers.

Note: Some symbols are not supported by all graphics output formats. If you select a symbol that is not supported by the output format, then a different symbol (one that is supported) is displayed instead.

5 Select whether to break the plot lines at missing values.

Note: The **Break the second plot line at missing values** check box is available only if you assign a variable to the **Vertical (Right)** role.

Selecting the Interpolation Method

In the selection pane under the Appearance heading, click **Interpolations** to access these options.

To select an interpolation method:

1 Select a variable. If you are creating a line plot or a two-dimensional scatter plot, the variables in the list are the variables that you assigned to the **Vertical** and **Vertical (Right)** roles. If you selected **Multiple line plots by group column** as the type of line plot, the list contains the unique values of the variable that you assigned to the **Group** role.

2 Select an interpolation method from the drop-down list. Depending on the interpolation method that you choose, you might need to specify additional settings. These interpolation methods are available:

- **LaGrange** draws a smooth curve through the data points. This method is used primarily when data consists of tabulated, precise values. A polynomial of the specified degree (1, 3, or 5) is fitted through the nearest 2, 4, or 6 points. In general, the first derivative is not continuous. If the values of the X variable are not strictly increasing, the corresponding parametric method is used.
- **Line** connects data points with straight lines. Points are connected in the order in which they occur in the input data.
- **Needle** draws a vertical line from each data point to a horizontal line at the 0 value on the vertical axis or the minimum value on the vertical axis if it is greater than 0. The horizontal line is drawn automatically.
- **Regression** creates a regression analysis plot. This interpolation method is available only if you assign numeric variables to the Horizontal and Vertical roles.
- **Scatter** suppresses any interpolation. A plot of the data points is created.
- **Smooth** fits a smooth line to the data by using a spline routine. This is a method for smoothing noisy data. The points on the plot do not necessarily fall on the line.
- **Spline** plots the line by using a spline routine.
- **STD** uses a solid line to connect the mean Y value with ± 1, 2, or 3 standard deviations for each X. By default, two standard deviations are used.
- **Step** plots the data by using a step function.

3 (Optional) Select the **Apply to all** check box to use the same interpolation method and property settings for all the variables in the list. This option is available only if there is more than one variable in the list.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. **Note:** This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. **Note:** If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.

Role Name	Description
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Adding Reference Lines

In the selection pane, these options appear under the Appearance > Axes heading.

Note: These options are not available for three-dimensional scatter plots.

Role Name	Description
Display Grid Lines	displays the grid lines. **Note:** This option is not available for bar charts.
Reference Lines	adds reference lines to the horizontal or vertical axis. For some plots, you can add reference lines to the Vertical Right axis. **Note:** For bar charts, reference lines are drawn for the response axis. For vertical bar charts, the reference lines appear as horizontal lines, and for horizontal bar charts, the reference lines appear as vertical lines. For three-dimensional plots, no reference lines are available.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. **Note:** If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. **Note:** This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. **Note:** The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task

- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

78

Scatter Plot Matrix

About the Scatter Plot Matrix Task

The Scatter Plot Matrix task creates a paneled graph of scatter plots for multiple combinations of variables. You can use options to overlay fit plots and ellipses on your scatter plots.

Requirement Name	Procedure and Product Names
SAS procedures used	SGSCATTER
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Scatter Plot Matrix: Creating a Matrix of Scatter Plots

Role Name	Description
Specifying a name for the task	identifies the task in the project tree and process flow. In the **Label** box, enter the name for the task.
Specifying the input data source	enables you to modify the input data source for the task. By default, the data source that you selected before opening the task is the input data source. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Task Filter** field.
Assigning the matrix and group variables	■ **Matrix Variables**specifies the variables to graph against each other to create the scatter plots. You must assign at least two variables to this role. ■ **Group Variable**specifies a classification variable to divide the values into groups. Assigning a variable to this role is not required.

Role Name	Description
Specifying the chart title	specifies the title that appears in the results. By default, the title for the results is Scatter Plot Matrix. However, you can customize this title. Click the down arrow to edit the text of the title. You can use macro variables in the title. The values for the macro variables are determined when you run the task.
Specifying a diagonal	adds graphs to the diagonal cells of the plot matrix. You can place histograms in the diagonal cells, and overlay those histograms with normal density curves or kernel density estimates. You can also specify whether the diagonal starts in the top left corner or the bottom left corner. If you do not select a graph to include in the diagonal, the diagonal cells contain the variable names.
Adding an ellipse	adds a confidence or prediction ellipse to each cell that contains a scatter plot. Both ellipse types assume bivariate normal distribution. You can choose from these types of ellipses: ■ **None** - no ellipse is added. ■ **Average** - adds a confidence ellipse for the population mean. ■ **Prediction** - adds a prediction ellipse for a new observation. You can also specify the confidence level for the ellipse.
Displaying a legend	specifies whether to display a legend for the scatter plot matrix. Legends are available only when you assign a group variable. When you assign a group variable, a legend is displayed by default. To remove the legend, clear the **Show legend** check box. Specify a title for the legend, and then select a position for the legend from the **Position** drop-down list. You can also specify whether to display a border around the legend.
Specifying a chart area	specifies the height and width of the chart. You can also specify whether to display data tips that appear when you move the pointer over a data point in the scatter plot. By default, the chart area is 800 x 800, and no data tips are displayed.
Specifying a footnote for the chart	specifies the footnote that appears in the results. By default, the date and time that you ran the task appear in the footnote of the chart. (In the **Chart Footnote** text box, SAS macro variables for the current date and time appear.) However, you can customize the footnote for your results. Click the down arrow to edit the text in the footnote. You can use macro variables in the footnote. To delete the footnote, delete all text in the text box and click **OK**. Then no footnote is included in the results.

79

Sort Data

About the Sort Data Task

Sort Data enables you to sort a data set in your project by any of its columns. You can sort the data by one or more columns, and you can select ascending or descending order for each column. The Sort Data task places the sorted data set as a new data item in your project.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SORT
Required SAS products	Base SAS
Recommended additional SAS products	None

Sort Data: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Sort by	When you assign one or more variables to this role, the table is grouped by the selected variable or variables. The order in which the variables appear within this role determines which variable is the primary sort key, which variable is the secondary sort key, and so on. The primary sort key is always the first variable that is listed within the **Sort by** role.
Columns to be dropped	When you assign one or more variables to this role, the output that is generated does not contain the specified variables. You can assign a maximum of (*n* - 1) variables to this role, where n is the total number of variables in the data set that are being sorted.

Sort Data: Setting Sort Data Options

In the selection pane, click **Options** to access these options.

Option Name	Description
Output order	You can determine the order of the data in the output by modifying the following options: ■ **Collating sequence** indicates what collating sequence to use when sorting character variables. You can use the collating sequence that is defined on the server, or you can select the ASCII or EBCDIC collating sequences, or you can use the Danish, Finnish, Norwegian, or Swedish national standards. You can also choose to sort character variables by using a custom-defined collating sequence that is defined by your installation site. Note: If you select **Custom**, your site must have a customized national sort sequence that has been defined. Your administrator can determine whether a customized national sort sequence is available. If no sort sequence is available, then the default collating sequence that is defined for the server is used. ■ **Maintain original data order within 'Sort by' groupings** groups the data according to the order that you set for the analysis roles. If this option is not selected, then the output data set is grouped in an undefined order within the sorted key groups.

Option Name	Description
Duplicate records	You can specify how you want to handle duplicate records in the output by using these options: ■ **Keep all records** keeps all of the records that are in the output data set, including all duplicates of records. ■ **Keep only the first records for each 'Sort by' group** eliminates any duplicate observations that have the same values for the **Sort by** group. If the **Maintain original data order within 'Sort by' groupings** option is selected, then the observation that is retained for each **Sort by** group will be the first one that is read from the original data set. However, if the **Maintain original data order within 'Sort by' groupings** option is not selected, the observation that is kept for each **Sort by** group cannot be predetermined. ■ **Do not keep adjacent duplicate records** compares each record to the previous record in the output data set. If an exact match is found, the duplicate record is not written to the output data set. Note: If you do not assign all variables to the **Sort by** role, some duplicate records might not be removed.
Advanced sorting options	You can set advanced options for how to sort the data in the output. You can modify the following options: ■ **Memory for sorting** specifies the maximum amount of memory that can be used for the Sort Data task. You can specify the amount of memory in bytes (B), kilobytes (KB), megabytes (MB), or gigabytes (GB). You can also specify to use all of the available memory or to use the default amount of memory that has been allocated on the server. ■ **Reduce temporary disk space requirements** indicates that during the Sort Data process, only the **Sort by** variables and the observation numbers are stored within temporary files, reducing the amount of storage necessary to perform the sort. In the final phase of the sort, the temporary file is used as an index to access the original data set and then to send the data to the results data set in the correctly sorted sequence. ■ **Force a sort of indexed data** indicates that you want to sort all data sets, including data sets that might already be sorted in the desired sequence or that contain a user-created index with keys that reflect those specified in the **Sort by** role. If you specify this option, the data set is sorted regardless of the current order of the data set or whether it contains an index.

Sort Data: Setting Results Options

In the selection pane, click **Results** to access these options.

You can designate a different name or storage location for the output data set. You can also save any duplicate records to a data set if you choose to eliminate any duplicate observations that have the same values for the **Sort by** group, or if you choose to keep only one of the records for any records whose values match exactly. For more information, see the **Duplicate records** options.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

80

Split Columns

About the Split Columns Task

The Split Columns task creates an output data set by splitting the unique combination of values of the selected columns in the input data set into multiple columns. You can use the output data set to individually analyze the columns that contain multiple rows of the input data set.

This task is useful when you have a data set in which one column contains multiple observations for different subgroups and you want to split the subgroup measures into separate columns.

For example, the following data set contains a single column with the monthly temperature readings for various locations across the state of North Carolina:

	Location	Month	Temperature
1	Asheville	January	46.9
2	Asheville	February	50.6
3	Asheville	March	59
4	Asheville	April	68
5	Asheville	May	74.9
6	Charlotte	January	51
7	Charlotte	February	52.8
8	Charlotte	March	61.7
9	Charlotte	April	70.5
10	Charlotte	May	79.6
11	Raleigh	January	51
12	Raleigh	February	54.1
13	Raleigh	March	62.1
14	Raleigh	April	71.6
15	Raleigh	May	79
16	Wilmington	January	56
17	Wilmington	February	59
18	Wilmington	March	65.8
19	Wilmington	April	74.4
20	Wilmington	May	81.1

The Split Columns task enables you to create the following data set by transposing the monthly temperature readings for each location into a column for each month:

	Location	January	February	March	April	May
1	Asheville	46.9	50.6	59	68	74.9
2	Charlotte	51	52.8	61.7	70.5	79.6
3	Raleigh	51	54.1	62.1	71.6	79
4	Wilmington	56	59	65.8	74.4	81.1

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	TRANSPOSE
Required SAS products	Base SAS
Recommended additional SAS products	SAS/STAT

Split Columns: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Columns to split	The column that you assign to this role contains the values that are transposed into the new columns. You must assign only one variable to this role.
Value identifier column	The column that you assign to this role provides the names for the new columns that contain the transposed values of **Columns to split**. The sequence of unique values in this column identifies which rows to transpose into the new columns. If you do not assign any grouping variables, this column must contain only one set of unique values, and the values of **Columns to split** are transposed into one row in the output data set. You must assign a column to the **Value identifier column** role. To select a grouping variable, assign a column to the **Group analysis by** role.
Value label column	The column that you assign to this role provides labels for the new columns that contain the transposed values of **Columns to split**. **Note:** To see the effect of this column, view the properties of the new columns in the output data set.
Group analysis by	Each variable that you assign to this role is used to segment the rows of **Column to split** into subgroups that are transposed into new columns. Each subgroup, defined by using a unique combination of the values of the grouping variables, becomes a row in the output data set. The number of rows that form a subgroup is determined by the number of unique values in the **Value identifier column**.

Split Columns: Setting Results Options

In the selection pane, click **Results** to access these options.

You can designate a different name for the output data set.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

You can add a prefix to the column names in the output data set. By default, the names of the columns are created from the unique values in the **Value identifier column**. To specify a prefix, select the **Use prefix** check box and enter the prefix in the text box. The default prefix is `Column`.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

81

Stack Columns

About the Stack Columns Task

The Stack Columns task creates an output data set by restructuring selected columns in the input data set so that these columns are transposed into observations. You can use the output data set to analyze values across multiple columns of the input data set. If you group the observations, the selected columns are divided into subgroups that are based on the unique combinations of the grouping values. Each subgroup forms a row of the output data set.

This task is useful when you have a data set in which each observation contains the same type of data in multiple columns and you want to analyze the data across several columns. For example, the following data set contains monthly temperature readings for various locations across the state of North Carolina:

	Location	January	February	March	April	May
1	Asheville	46.9	50.6	59	68	74.9
2	Charlotte	51	52.8	61.7	70.5	79.6
3	Raleigh	51	54.1	62.1	71.6	79
4	Wilmington	56	59	65.8	74.4	81.1

The structure of the data enables you to compute summary statistics for each month based on all the locations in the state. However, you cannot calculate summary statistics for a specific location, because the statistics must be computed across an observation.

To compute statistics for all the temperatures or for each location, you can stack the columns with the monthly temperature readings by location to create the following output data set:

	Location	Month	Temperature
1	Asheville	January	46.9
2	Asheville	February	50.6
3	Asheville	March	59
4	Asheville	April	68
5	Asheville	May	74.9
6	Charlotte	January	51
7	Charlotte	February	52.8
8	Charlotte	March	61.7
9	Charlotte	April	70.5
10	Charlotte	May	79.6
11	Raleigh	January	51
12	Raleigh	February	54.1
13	Raleigh	March	62.1
14	Raleigh	April	71.6
15	Raleigh	May	79
16	Wilmington	January	56
17	Wilmington	February	59
18	Wilmington	March	65.8
19	Wilmington	April	74.4
20	Wilmington	May	81.1

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	TRANSPOSE
Required SAS products	Base SAS
Recommended additional SAS products	SAS/STAT

Stack Columns: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Columns to stack	Each column that you assign to this role becomes one or more rows of the output data set. When you select one or more grouping variables, the grouping variables are used to segment each column into subcolumns, each of which is turned into a row. In this case, a column is transposed to the number of rows that is equal to the number of groups that are defined by the grouping variables. You must assign at least one column to the **Columns to stack** role. To select a grouping variable, assign a column to the **Group analysis by** role. **Note:** If you do not assign any grouping variables, then the input data set must contain only one observation and each column becomes one row of the output data set.
Group analysis by	Each variable that you assign to this role is used to segment the columns to stack into subgroups that are transposed separately. Each subgroup, defined by using a unique combination of the values of the grouping variables, becomes a row of the output data set. Each unique combination of the values of the grouping variable occurs only once.

Stack Columns: Setting Column Naming Options

In the selection pane, click **Column Naming** to specify the names and labels for the new columns in the output data set.

Role Name	Description
Specifying the name and label for the New value column	A new column is created by transposing the values of an observation in the input data set into values of a variable in the output data set. To specify a heading for the output column that contains the transposed values, enter a heading in the **Name** box in the **New value column** area. Initially, the **Name** box contains the default heading `StackedValues`. To specify a description for the column that is stored with the output data set, enter a label in the **Label** box in the **New value column** area. The character description string can be up to 256 characters long. Leave this box blank to omit the label.
Specifying the name and label for the Derived column	Each row of the output data set includes the name of the variable in the input data set that is transposed to create the current observation. To specify a heading for the output column that contains these variable names, enter a heading in the **Name** box in the **Derived column** area. Initially, the **Name** box contains the default heading `ValueSource`. To specify a description for the column that is stored with the output data set, enter a label in the **Label** box in the **Derived column** area. The character description string can be up to 256 characters long. Leave this box blank to omit the label.
Specifying the name and label for the Description column	Each row of the output data set includes the label of the variable in the input data set to which the values in that output row belong. To specify a heading for the output column that contains these variable names, enter a heading in the **Name** box in the **Description column** area. Initially, the **Name** box contains the default heading `ValueDescription`. To specify a description for the column that is stored with the output data set, enter a label in the **Label** box in the **Description column** area. The character description string can be up to 256 characters long. Leave this box blank to omit the label. **Note:** To see the effect of specifying the label for a column, view the properties of the new column in the output data set.

Stack Columns: Setting Results Options

In the selection pane, click **Results** to access these options.

You can designate a different name for the output data set.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

82

Standardize Data

About the Standardize Data Task

The Standardize Data task standardizes variables in a SAS data set to a given mean and standard deviation, and it creates a new SAS data set that contains the standardized values.

For example, you can use this task to standardize test scores by using a specified mean and standard deviation.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	STANDARD
Required SAS products	Base SAS
Recommended additional SAS products	None

Standardize Data: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	performs standardizations on each variable that you assign to this role. You must assign at least one variable to this role.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies the values to use to calculate a weighted mean and a weighted variance for the analysis variables. You can assign a maximum of one variable to this role.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group the table by a variable that you have already assigned to another role.

Standardize Data: Setting Standardization Options

In the selection pane, click **Standardize** to access these options.

Option Name	Description
Specifying a new mean and standard deviaion	By default, this task standardizes the data to a mean of 0 and a standard deviation of 1. You can specify a new mean and standard deviation for the variable or variables that you are analyzing, or you can choose to retain the current mean or standard deviation. **Note:** The same mean and standard deviation are applied to all analysis variables. You can choose to retain each variable's mean or standard deviation by clearing the **New mean** or **New standard deviation** check boxes.
Handling missing values	By default, missing values in an input analysis variable are excluded from the standardization process. You can replace the input missing values with the mean of the nonmissing values for that variable by selecting the **Replace missing values with column mean value** check box.

Standardize Data: Setting Results Options

In the selection pane, click **Results** to access these options.

By default, the **Add new standardize variables** option is selected so that the standardized variables are included with the original variables in the output data. To replace the original variables with the standardized variables, select **Replace existing variables**.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

83

Summary Statistics

About the Summary Statistics Task

The Summary Statistics task provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations. You can also summarize your data in a graphical display, such as a histogram.

For example, you could use this task to create a report on the number of new sales, arranged by product type and country.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	MEANS, UNIVARIATE
Required SAS products	Base SAS
Recommended additional SAS products	SAS/GRAPH

Summary Statistics: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the variable on which you want to have statistics generated. You must assign at least one variable to this role.
Classification variables	specifies variables to use to divide the input data into categories or subgroups. These variables can be character or numeric. The statistics are calculated on all selected analysis variables for each unique combination of classification variables. You can set the **Allow multilabel formats** option so that the primary and secondary format labels for a given range or overlapping ranges are used to create subgroup combinations when a multilabel format is assigned to this variable.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Relative weight	specifies variable to use to calculate the weighted means, variances, and sums.
Copy variables	specifies the variables to include in the output data set. You can use the **Value to copy for Copy Variables role** option to copy either the maximum or minimum value that the variable has in the corresponding observations of the input data set to the output data set.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables.

Summary Statistics: Selecting Summary Statistics

Selecting Basic Summary Statistics

In the selection pane under the Statistics heading, click **Basic** to access these options.

By default, the Summary Statistics report includes the following statistics: mean, standard deviation, minimum, maximum, and number of observations. You can control the statistics that are computed, the number of decimal places that are displayed in the report, and the method that is used to compute the quantile statistics.

Option Name	Description
Selecting basic statistics	specifies the basic statistics to include in the analysis.
Maximum decimal places	specifies the maximum number of decimal places for the calculated statistics. By default, a statistic is displayed by using the best fit, which is usually seven decimal places.

Option Name	Description
Specifying a divisor for standard deviation and variance	specifies the divisor that is used in the calculation of standard deviations and variances. By default, the divisor is the degrees of freedom.

Selecting Percentile Summary Statistics

In the selection pane under the Statistics heading, click **Percentiles** to access these options.

Option Name	Description
Percentile statistics	specifies the percentiles to compute. You can also choose to calculate the lower quartile (the 25th percentile), the median (the 50th percentile), and the upper quartile (the 75th percentile). By default, no percentiles are selected.
Quantile method	specifies the method that is used to compute the quantiles, median, and percentiles. You can select from these two methods: ■ **Order statistics** to use a method that reads all of the data into memory and sorts it by the unique values. ■ **Piecewise-parabolic algorithm** to use a less memory intensive method that approximates the quantile.

Selecting Additional Summary Statistics

In the selection pane under the Statistics heading, click **Additional** to access these options.

You can also choose to calculate these statistics:

- **Confidence limits of the mean**
- **t statistic and Prob > |t|**
- **Coefficient of variation**
- **Corrected sum of squares**
- **Uncorrected sum of squares**

You can also calculate the confidence level for the confidence limits of the mean.

Summary Statistics: Setting the Plot Options

In the selection pane, click **Plots** to access these options.

Note: If you assigned a variable to the **Classification variables** role, the plots that are generated are only for the *n*-way combination of the classification variables.

Plot Name	Description
Histogram	A histogram is a graph that is used to determine the distribution of a set of data. You can specify midpoints for the histogram. The options that are available from the drop-down list depend on whether you are creating a histogram or a comparative histogram. If you assign a variable to the **Classification variables** role, you are creating a comparative histogram. For comparative histograms, the following options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Key** determines the midpoints for the data in the key cell. By default, the levels for the classification variable are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first classification variable. If you specified only one classification variable, then the key cell is the level that occurs first in this order. If you specified two classification variables, then the key cell is the combination of levels of variable 1 and variable 2 that occurs first in this order. Thus, the choice of the key cell determines the uniform horizontal axis that is used for all cells. The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. ■ **Uniform** determines the number of midpoints based on the total sample size. If you do not assign a variable to the **Classification variables** role, you are creating a histogram. For histograms, these options are available: ■ **Specify midpoints** enables you to specify the starting and ending midpoints. You can also specify the step interval to use to calculate the other midpoints. The midpoint is the middle value of the range of values represented by each bar. ■ **Specify number of bins** enables you to specify the number of bins (also called histogram intervals) for the data. By default, the number of bins that are displayed is based on the range of the data. However, you can specify the maximum number of bins to display.
Box and whisker	Box plots show a measure of central location (the median), two measures of dispersion (the range and interquartile range), the skewness (from the orientation of the median relative to the quartiles), and potential outliers. Box plots are especially useful in comparing two or more sets of data.

Summary Statistics: Setting Results Options

In the selection pane, click **Results** to access these options.

Option Name	Description
Save statistics to data set	specifies whether to permanently save the statistics to a data table. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Option Name	Description
Selecting the statistics to include in the results	specifies the statistics to include in the results. You can choose from these options: ■ **Value to copy for Copy Variables role** - If you assign a variable to the Copy variables role, then you can choose to copy either the maximum or minimum value that the variable has in the corresponding observations of the input data set to the output data set. ■ **Combinations of classification variables** - The **Classification variable** role enables you to group the summary statistics for the classification variables into tables. You can create up to n-way tables, where n is the number of classification variables. The groups are determined by the unique combinations of the classification variables. The order of the variables in the table depends on the order of the variables in the **Classification variables** role. You can use the following options to subgroup the output data: □ **N-way only** creates a single table that shows the summary statistics for the n-way combination of the classification variables. For example, suppose that you want to compare the sales that result from two different marketing campaigns. You assign the variable Profit to the **Analysis variables** role and the variables Month, Marketing Campaign, and Region to the **Classification variables** role. Then selecting the **N-way only** option results in a table of Month*Campaign*Region. □ **All ways** creates a separate table of summary statistics for each unique combination of the classification variables. For example, using the variables Month, Campaign, and Region and then selecting the **All ways** option results in the following tables: ▪ a table of overall sales. This is called a zero-way table, and the classification variables are not included. ▪ a one-way table that displays the summary statistics for each month (for example,, May, June, and July) ▪ a one-way table that displays the summary statistics for each region (for example,, north, south, east, and west) ▪ a one-way table that displays the summary statistics for each type of marketing campaign (for example,, TV commercial, magazine ad) ▪ a two-way table that displays the summary statistics for Month*Region ▪ a two-way table that displays the summary statistics for Month*Campaign ▪ a two-way table that displays the summary statistics for Region*Campaign ▪ a three-way table that displays the summary statistics for Month*Campaign*Region □ **Specify ways** creates a separate table of summary statistics for each unique combination of the classification variables that you specify. For example, if you enter `1`, `2`, or `3` in the **Specify ways** box, the results contain all the possible one-way, two-way, or three-way tables. If you specify `0` in this box, then the results contain a zero-way table, which does not include the classification variables.
Show Analysis labels	displays the labels for the analysis variables.
Suppress all displayed output	does not display the results. If you choose to suppress the displayed output, you must save your results to an output data set.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

84

Summary Statistics Wizard

About the Summary Statistics Wizard

The Summary Statistics wizard helps you create tables of descriptive statistics as well as histograms and box plots, with optional grouping variables.

Summary Statistics Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to assign variables to roles.

Summary Statistics Wizard: Assign Variables to Roles

1 Before you can assign variables to roles, you must select the data.

2 Assign at least one variable to the **Summary Statistics of (Analysis variable)** role. The variables that you assign to this role are the numeric variables on which you want to have statistics generated.

   Note: You must assign at least one variable to this role. All other roles are optional.

3 Assign at least one variable to the **For each value of (Classification variable)** role. The variables that you assign to this role are character or discrete numeric variables that are used to divide the input data into categories or subgroups. The statistics are calculated on all selected analysis variables for each unique combination of classification variables.

4 Assign at least one variable to the **Separate tables for values of (Group variable)** role. The variables that you assign to this role are used to compute separate statistics for each distinct value or combination of values of the **Group analysis by** variables. The data is automatically sorted by the variables in this role before the statistics are computed. To select the sort order for each variable, use the up and down arrows.

5 (Optional) Click **Advanced** to assign variables to these roles:

- **Specify frequency by (Frequency count)** - When you assign a variable to this role, each observation in the table is assumed to represent n observations, where n is the value of the frequency count for that row. Statistics are calculated accordingly. You can assign a maximum of one variable to this role.
- **Weight analysis by (Relative weight)** - If you assign a variable to this role, the value of the variable for each observation is used to calculate weighted means, variances, and sums. You can assign a maximum of one variable to this role.
- **Include labels of (Copy variable)** - The variables that you assign to this role are included as additional variables in the output data set.

6 Click **Next** to specify the statistics and results.

Summary Statistics Wizard: Specify Report Statistics and Output Results

1 Before you can specify the report statistics and output results, you must assign variables to roles.

2 Select the statistics to include in the results. By default, the mean, standard deviation, minimum, maximum, and number of observations are included. You can include additional statistics by using the options on these tabs:

Basic
includes the descriptive options. You can also select the maximum number decimal places for the calculated statistics. By default, a statistic is displayed by using the best fit, which is usually seven decimal places.

Percentile
specifies the percentiles to compute. You can also choose to calculate the lower quartile (the 25th percentile), the median (the 50th percentile), and the upper quartile (the 75th percentile). By default, no percentiles are selected.

Use the **Quantile method** option to select the method that is used to compute the quantiles, median, and percentiles. You can select from these two methods:

- **Order statistics** to use a method that reads all of the data into memory and sorts it by the unique values.
- **Piecewise-parabolic algorithm** to use a less memory intensive method that approximates the quantile.

Additional
includes these statistics:

- **Confidence limits of the mean**
- **t statistic and Prob > |t|**
- **Coefficient of variation**
- **Corrected sum of squares**
- **Uncorrected sum of squares**

You can also calculate the confidence level for the confidence limits of the mean.

3 Select the **Show statistics** option to display the results. You must choose either to show the statistics in the results or to save them permanently to a data table, or both.

4 Select the **Histogram** or **Box and whisker** check box to select the type of plot that you want to generate. You can choose one or both. By default, no plots are selected.

Note: If you assigned a variable to the **Classification variables** role, the plots that are generated are only for the *n*-way combination of the classification variables.

5 Select the **Save statistics to data set** check box to save the output. You must choose either to show the statistics in the results or to save them permanently to a data table, or both.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

6 Click **Next** to specify the titles and footnotes.

Summary Statistics Wizard: Specify the Title and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1 Before you can specify the title and footnotes, you must specify report statistics and output results.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

Note: Some titles might not be available, depending on the plots that you selected to include in the results.

3 Click **Finish** to run the task.

85

Summary Tables

About the Summary Tables Task

The Summary Tables task displays descriptive statistics in tabular format, using some or all of the variables in a data set. You can create a variety of tables, ranging from simple to highly customized.

For example, suppose that you have a SAS data set that contains data on sales of a product by two types of customers, residential and business, in individual states in the northeast and west regions of the United States. You can use the Summary Tables task to sum the sales for each customer type in each state in a geographic region.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	TABULATE
Required SAS products	Base SAS
Recommended additional SAS products	None

Summary Tables: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Analysis variables	specifies the variables for the analysis. You must assign at least one variable to this role or the **Classification variables** role.
Classification variables	specifies the character or discrete numeric variables to use to divide the input data into categories or subgroups. The statistics are calculated on all selected analysis variables for each unique combination of classification variables. You can set the **Multilabel formats** option so that the primary and secondary format labels for a given range or overlapping ranges are used to create subgroup combinations when a multilabel format is assigned to this variable.
Pages	generates a separate summary table for each distinct value of this variable.
Relative weight	specifies variable to use to calculate the weighted means, variances, and sums.
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.

Summary Tables: Setting Up the Summary Table

In the selection pane, click **Summary Tables** to access these options.

To set up a summary table, you must decide what information you want and how you want it to appear. The Summary Tables task is very flexible; you can arrange classification and analysis variables in any way that you want.

To use a variable:

1. Select the variable in the **Available variables** box.
2. Drag the variable to a cell in either the top or left portion of the **Preview** box. If you drop the new variable onto an existing variable, the new variable replaces the old one. To add the new variable to the existing variables, drop it onto a cell border. The border turns blue when the mouse pointer is in the correct position.

   Note: By default, the statistic applied to the analysis variables that you use is number of observations (N). You can add or replace statistics using the same methods that are described above.

Summary Tables: Setting Additional Options

In the selection pane, click **Summary Tables** to access these options.

Right-click on the preview table to set options that customize the appearance of the summary table. With the exception of **Remove Cells**, the items on the pop-up menu each open a window that enables you to set various options.

Plot Name	Description
Remove cells	This item is active when you right-click on an analysis variable, a classification variable, or a statistic in the preview table. Use this item to remove the variable or statistic from the summary table.
Table Properties	This item is always active. Options that are set in the Table Properties dialog box are applied to the entire summary table. ■ **General** tab: Specify a label for missing values. The default is a period ('.'). Use the check boxes to include missing values as a legitimate value for classification variables and to suppress the headings (labels) for the row classification variables. ■ **Format** tab: Assign a standard SAS format or user-written format to all of the data values in the summary table. Select the format, field width, and number of decimal places. ■ **Font** tab: Specify the font, size, style, foreground color, background color, justification, and vertical alignment for all text in the table.
Box Area Properties	This item is always active. Use this item to set options for the upper left corner of the table. These options override any options that are selected in the Table Properties dialog box. ■ **General** tab: Enter the text to be used in the Box Area. If you have assigned one or more variables to the Pages role, you can choose to show the page headings in the Box Area. Also, you can choose whether to wrap the text in the Box Area or truncate it if the text is too wide for the Box Area. ■ **Format** tab: This tab is not applicable to the Box Area. ■ **Font** tab: Specify the font, size, style, foreground color, background color, justification, and vertical alignment for the text in the Box Area.
Heading Properties	You must right-click on a column or row heading to activate this item. Use this item to set options for individual column or row headings. These options override any options that are selected in the Table Properties dialog box. ■ **General** tab: Enter a label for the row or column heading. Also, choose whether to wrap the text in the heading or truncate it if it is too wide for the heading space. ■ **Format** tab: This tab is not applicable to classification variable values. To assign a heading format to a classification variable, click **Data** in the selection pane, and select the classification variable. Select the Heading format option to assign the format. ■ **Font** tab: Specify the font, size, style, foreground color, background color, justification, and vertical alignment for the heading text.
Class Level Properties	This item is active only when you right-click on a classification variable in the preview table. Use this item to set options for the classification variable values. ■ **General** tab: Choose whether to wrap the classification variable value text. ■ **Format** tab: This tab is not applicable to classification variable values. To assign a heading format to a classification variable, click **Data** in the selection pane, and select the classification variable. Select the **Heading format** option to assign the format. ■ **Font** tab: Specify the font, size, style, foreground color, background color, justification, and vertical alignment for the classification variable value text.

Plot Name	Description
Data Value Properties	This item is active when you right-click on any variable or statistic in the preview table. Use this item to set options for the data values in the selected row or column of the summary table. These options override any options that are selected in the Table Properties dialog box. ■ **General** tab: Choose whether to wrap the data value text. ■ **Format** tab: Assign a standard SAS format or user-written format to all of the data values in the selected row or column of the summary table. Select the format, field width, and number of decimal places. ■ **Font** tab: Specify the font, size, style, foreground color, background color, justification, and vertical alignment for the data value text.

Summary Tables: Setting Results Options

In the selection pane, click **Results** to access these options.

To save your results in an output data set, select the **Save results to a data set** check box.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

86

Summary Tables Wizard

About the Summary Tables Wizard

The Summary Tables wizard helps you create tables in a variety of configurations including crosstabulation tables, with options for summary statistics.

Summary Tables Wizard: Select the Data

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

Click **Next** to assign variables and statistics.

Summary Tables Wizard: Assign Variables to Roles

1 Before you can assign variables to roles, you must select the data.

2 To add a variable to the **Analysis variables** box, click **Add**. From the pop-up menu, select the variable to add to this role. The variables that you assign to this role are the numeric variables for the calculations.

3 In the **Analysis variables** box, select the statistic that you want to calculate for each analysis variable. By default, the sum statistic is calculated. To calculate a different statistic, select the cell in the **Statistic** column for the statistic that you want to change. From the drop-down list, select the statistic to calculate.

Note: You cannot assign an analysis variable if the data does not contain any numeric variables. When this happens, the frequency statistic is used in place of an analysis variable.

4 From the **Analysis variables labels** drop-down list, select the location of the labels for the analysis variables. The available options include **in columns**, **in rows**, **in pages**, or **hidden**.

5 From the **Statistics labels** drop-down list, select the location of the labels for the statistics. The available options include **in columns**, **in rows**, **in pages**, or **hidden**.

6 Click **Next** to specify the statistics and results.

Summary Tables Wizard: Select the Classification Variables

1 Before you can select the classification variables, you can assign the analysis variables and statistics.

2 To add a classification variable to the columns, rows, or pages, click **Add** for that box. From the pop-up menu, select the variable to add. The variables that you assign as classification variables are character or discrete numeric variables that are used to divide the input data into categories. The statistics will be calculated on all selected analysis variables for each unique combination of classification variables. A separate summary table is generated for each variable that you assign to the **Pages** box.

3 Click **More Options** to customize the table structure. In the Table Structure dialog box, you can choose to hide the headings for the columns, rows, and pages. You can also choose to nest or concatenate the columns, rows, and tables.

4 Click **Next** to specify the titles and footnotes.

Summary Tables Wizard: Specify the Totals

1 Before you can specify the totals, you must specify the classification variables.

2 From the **Columns**, **Rows**, and **Pages** drop-down lists, select whether to display the grand total for each column, row, or page. To display the total, select **Grand total only**. If you do not want to see the total, select **None**. For pages, you can also specify whether to display the page totals first or last on the page.

Note: If you did not specify a classification variable for the columns, rows, or pages, then the corresponding option is not available.

3 In the **Label for totals** box, enter the label for the total.

4 Click **Next** to specify the titles and footnotes.

Summary Tables Wizard: Select Additional Output

1 Before you can select the additional output, you can specify the totals.

2 Select the **Save results to a data set** check box.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

3 Click **Next** to specify the titles and footnotes.

Summary Tables Wizard: Specify the Title and Footnotes

To accept the default title and footnote text for the results, you do not need to do anything.

1 Before you can specify the title and footnotes, you can specify whether to save the results to an output data set.

2 Edit the text of the title or footnote in the text box. You can use macro variables in the titles and footnotes.

Note: Some titles might not be available, depending on the plots that you selected to include in the results.

3 Click **Finish** to run the task.

87

Surface Plot

About the Surface Plot Task

The Surface Plot task creates three-dimensional wireframe plots, three-dimensional smooth plots, or three-dimensional gradient plots that show the mathematical relationships between three numeric variables. In the three-dimensional plot, one vertical variable (z) is plotted for a position on a plane that is specified by two horizontal variables (x and y). The coordinates of each point correspond to the values of three numeric variable values in an observation of the input data set. The observation can contain values in the form z = f(x,y) or independent values such as the altitude at a given longitude and latitude.

For more information, see “Selecting Graphics Output Format” on page 3.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	G3D
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Surface Plot: Selecting a Plot Type

In the selection pane, click **Surface Plot** to access these options.

Role Name	Description
3D Wire Frame Surface Plot	creates a three-dimensional plot in which the wireframe grid displays the mathematical relationship between three variables.
3D Smooth Surface Plot	creates a three-dimensional plot in which the surface is all one color.
3D Surface Plot with Banding	creates a three-dimensional plot that uses color to indicate a change in value. The regions are clearly defined.
3D Smooth Gradient Surface Plot	creates a three-dimensional plot that uses color to indicate a change in value. The regions are not as clearly defined as in the gradient surface plot with banding.

How the surface plot displays depends on the graphic output format that you selected.

ActiveX and ActiveX image
: renders all surface plot types.

Java and Java image
: renders the three-dimensional wireframe surface plot. The three-dimensional smooth surface plot, the three-dimensional gradient surface plot with banding, and the three-dimensional smooth gradient surface plot are rendered as three-dimensional smooth gradient surface plots.

GIF, JPEG, PNG, and SAS EMF
: renders all surface plot types as three-dimensional wireframe surface plots.

Surface Plot: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Horizontal	The column that you assign to this role is the horizontal or X axis variable for the chart.
Vertical	The column that you assign to this role is the vertical or Y axis variable for the chart.
Depth	The column that you assign to this role is the third-dimension or Z-axis variable for the chart. **Note:** This option is available only for three-dimensional plots.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value.

Surface Plot: Setting Appearance Options

Setting Surface Plot Options

In the selection pane under the Appearance heading, click **Surface** to access these options.

Select a color for the top (facing the positive-Z direction) and bottom (facing the negative-Z direction) of the chart.

Select whether to draw the sides of the chart, creating a solid.

Use the sliders or the **Rotate** and **Tilt** controls to set the viewing angle. By default, the rotate angle is 70, and the tilt angle is 70. Click **Reset** to return the viewing angle to the default setting. Note

Note: If you clear the **Specify rotate and tilt values** check box, the rotate and tilt of the graph are determined by the graph output format. By default, this option is selected, and you can specify the viewing angle.

For the 3-D banded gradient and 3-D smooth gradient plots, you can select whether to draw a contour of the data along with the surface. You can display the contour above or below the surface.

Setting Axis Options

In the selection pane, these options appear under the Appearance > Axes heading.

Note: The axis options are not applied if you are creating a surface plot by using GIF as the graphical output format. To apply the axis options in a surface plot, use ActiveX as the graphical output format.

Role Name	Description
Turn off Axes and Ticks	suppresses the axes and tick marks.

Role Name	Description
Axis	customizes the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter, three-dimensional needle, and three-dimensional surface plots, you can specify these options for the Depth axis. ■ For group plots or multiple measures bar charts, you can specify these options for the Group axis. To customize the appearance of an axis: ■ For horizontal, vertical, and depth axes, specify the color, width, and style of the line for the axis. ■ To reverse the order of the values for the horizontal or vertical axis, select the **Reverse Axis** option. If you select this check box for the horizontal axis, then the order of the X values is reversed. If you select this check box for the vertical axis, then the order of the Y values is reversed. Note: This option is available only for area plots, box plots, bubble plots, contour plots, line plots, and two-dimensional scatter plots. ■ To customize the label of an axis, click the **Label** tab. Type the text for your custom label in the **Label** box. Use the **Label rotation** drop-down list to set how the label should be displayed in relation to the axis. You can also specify the type, size, and color of the font. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Note: If you are creating a multiple measures bar chart, then the default label for the response axis is the statistic used to calculate the length of each bar. ■ To customize the values of an axis, click the **Values** tab. You can specify the type, size, and color of the font for the values. You can choose whether to underline, boldface, or italicize the text. Use the buttons to change the justification. Use the **Values rotation** drop-down list to specify how the values should be displayed in relation to the axis.
Major tick marks	specifies the major tick marks for the horizontal or vertical axis. For some plots, you can also set options for the Vertical Right axis. For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis. Select **Show major *axis-name* ticks** and specify how to determine the number of ticks to use. ■ To let the software determine the number and frequency of the major tick marks, select **Automatic**. If you are creating an area plot, bubble plot, line plot, or two-dimensional scatter plot, select the **Begin at zero** check box to force the tick mark values to start at 0. ■ To specify the number of tick marks to use, select **Use**. Enter the desired number in the text box. ■ To specify a logarithmic tick pattern, select **Log**. Select a base and a power, and select whether to arrange the tick marks by values of the base or values of the power. ■ To specify particular locations for tick marks, select **Specify**. Type a tick mark value in the text box and click **Add**. To change a value, select it from the list and click **Edit**. After you finish making your changes, click **Save**.

Role Name	Description
Minor Ticks	specify minor tick marks for the horizontal or vertical axis. ■ For some plots, you can also set options for the Vertical Right axis. ■ For the three-dimensional scatter plots, three-dimensional scatter plots with needles, and three-dimensional surface plots, you can create tick marks for the Depth axis. ■ For bar charts, you can specify tick marks only for the response axis. For example, if you are creating a vertical bar chart, then you can specify tick marks for the vertical axis but not the horizontal axis.

Customizing the Chart Area

In the selection pane under the Appearance heading, click **Chart Area** to access these options.

Role Name	Description
Specify custom chart size	specifies the the image size. For reference, your current screen size is given.
Chart background color	specifies the background color. Note: If you are creating a map chart, then you choose the background color of the map from the **Map background color** drop-down list.
Draw frame around plot area	includes a frame around the chart area. Choose a color for the plot area from the **Plot area background color** drop-down list. Note: This option is not available for donut charts, map charts, pie charts, three-dimensional grouped bar charts, three-dimensional scatter plots, or surface plots.
Displaying chart tips	specifies whether to display a chart tip. The chart tip applies to the entire graph. It is not for the individual components of the graph, such as bars or lines. Note: The chart tip is not visible for map charts or if you selected ActiveX or Java as the format of the graph output.

Smoothing the Data

In the selection pane under the Appearance heading, click **Grid Data** to access these options.

You might want to create a rectangular grid of interpolated or smoothed values in these cases:

- where the data points in your input table are irregularly spaced
- where the input table has an insufficient number of rows to produce a three-dimensional surface plot
- where the data in your input table is noisy

By default, no interpolation method is used. However, you can select from these interpolation methods:

Join
uses a linear interpolation within a set of triangular regions that are formed from the input data set. This interpolation method creates values in the range of the initial values of the vertical variable, but the resulting interpolated surface might not be smooth.

No scale
specifies that the X and Y variables not be scaled to the same range before interpolation. By default, the interpolation is performed after both variables are similarly scaled because the interpolation methods assume that the scales of X and Y are comparable.

Partial
specifies the use of a spline to estimate the derivatives for the biquintic polynomial interpolation. A bivariate spline is fit to the nearest neighbors and used to estimate the needed derivatives.

Spline
specifies the use of a bivariate spline to interpolate. This option results in the use of an order n3 algorithm, where n is the number of input data points. This method can be time-consuming.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

88

t Test

About the t Test Task

The *t* Test task performs *t* tests for one sample, two samples, and paired observations. The one-sample *t* test compares the mean of the sample to a given number. The two-sample *t* test compares the mean of the first sample minus the mean of the second sample to a given number. The paired-observations *t* test compares the mean of the differences in the observations to a given number.

Sometimes it is not feasible to assume that two groups of data are independent, and a natural pairing of the data exists. In this case, it is advantageous to use a paired *t* test. For example, suppose that a stimulus is being examined to determine its effect on systolic blood pressure. Twelve men participate in the study, and their systolic blood pressure is measured both before and after the stimulus is applied. You can use the *t* Test task to determine whether the mean change in systolic blood pressure is significantly different from zero.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	TTEST
Required SAS products	Base SAS, SAS/STAT
Recommended additional SAS products	SAS/GRAPH

t Test: Choosing the Type of t Test

In the selection pane, click **t Test Type** to access these options.

Requirement Name	Procedure and Product Names
Two Sample	compares the mean of the first sample minus the mean of the second sample to a given number.
Paired	compares the mean of the differences in the observations to a given number.
One Sample	compares the mean of the sample to a given number.

t Test: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Classification variables	Available for two-sample *t* tests. Assign a classification variable to this role for analysis of two independent samples. The variable must have two levels only, one for each sample. You can assign a maximum of one variable to this role.
Analysis variables	Available for one-sample and two-sample *t* tests. t tests are performed on each variable in this list. You must assign at least one variable to this role.
Paired variables	Available for paired *t* tests. To perform a paired *t* test, assign the two variables to be analyzed to this role. Because paired *t* tests are performed by subtracting each value of the second variable from the corresponding value of the first, the order of the variables matters. **Note:** You must assign exactly two variables to this role.
Group analysis by	When you assign one or more variables to this role, the table is sorted by the selected variable or variables, and separate t tests are performed for each group. **Note:** You cannot group *t* tests by a variable that you have already selected for **Analysis variables** or **Paired variables**.
Frequency count	When you assign a variable to this role, each row in the table is assumed to represent n observations, where n is the value of the frequency count for that row. You can assign a maximum of one variable to this role.
Relative weight	When you assign a variable to this role, the value of the variable in each observation is used to compute weighted statistics. You can assign a maximum of one variable to this role. **Note:** If the input data set for this task contains only summary statistics, then the **Relative weight** role is not available. The weight option has no effect when the input data set contains summary statistics.

t Test: Setting Analysis Options

In the selection pane, click **Analysis** to access these options.

Option Name	Description
Specifying a null hypothesis	By default the null hypothesis has a value of 0. To specify a new null hypothesis value, enter the new value in the **H_0=** box.
Setting standard deviation confidence interval options	Set confidence interval options as follows: ■ The **Equal tailed** check box is selected by default. The equal-tailed confidence interval puts an equal amount of area in each tail of the chi-square distribution. This option requests that an equal-tailed confidence interval be displayed. ■ Select the **UMPU** check box to yield a confidence interval that is derived from the uniformly most powerful unbiased test. Note: To request that no confidence interval be displayed, clear both the **Equal tailed** and **UMPU** check boxes.

t Test: Generating Plots

In the selection pane, click **Plots** to access these options.

Plot Name	Description
Summary Plot	generates a summary plot. The contents of the summary plot depends on the type of *t* test that you chose to perform. ■ For a one-sample *t* test, the summary plot contains a histogram with overlaid normal and kernel densities, a box plot, and a confidence interval band. ■ For a two-sample *t* test, the summary plot contains comparative histograms with overlaid densities and box plots. ■ For a paired *t* test, the summary plot contains a histogram, densities, a box plot, and confidence interval of the difference or ratio.
Histogram	generates a histogram or comparative histograms with overlaid normal and kernel densities. For one-sample *t* tests and paired *t* tests, the histogram and densities are based on the test criterion (which is the mean difference or ratio for a paired design). For two-sample *t* tests, comparative histograms (one for each class) are shown.
Box plot	generates a box plot or comparative box plots. The box is drawn from the 25th percentile (lower quartile) to the 75th percentile (upper quartile). The vertical line inside the box shows the location of the median. For one-sample and paired *t* tests, a confidence interval for the mean is shown as a band in the background. For two-sample *t* tests, comparative box plots (one for each class) are shown.
Confidence interval plot	generates plots of the confidence interval for the means.

Plot Name	Description
Normal quantile-quantile (Q-Q) plot	generates a normal quantile-quantile (Q-Q) plot. For two-sample *t* tests, separate plots are shown for each class in a single panel.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

89

Table Analysis

About the Table Analysis Task

The Table Analysis task enables you to generate crosstabulation tables, also known as contingency tables, from your data.

For example, suppose that you want to compare the probability of coronary heart disease for two different types of diet. In a SAS data set, you have data for a case-control study of a high-fat diet and the risk of coronary heart disease. The data is recorded as cell counts, where the variable Count contains the frequencies for each diet and risk combination. Using the Table Analysis task, you can create a 2x2 contingency table that displays the frequency values for each diet (high fat and low fat) and risk (high risk and low risk) combination.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	FREQ
Required SAS products	Base SAS
Recommended additional SAS products	None

Table Analysis: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Frequency count	specifies the variable that contains the frequency count for each row in the table. Each role is assumed to represent *n* observations, where *n* is the value of the frequency count for that row.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** You cannot group analyses by a variable that you assigned to other roles.
Table variables	uses these variables to create two-way to n-way frequency and crosstabulation tables. Statistics are computed for the table variable(s). You must assign at least two variables to this role.

Table Analysis: Defining Tables

In the selection pane, click **Tables** to access these options.

To generate a new table:

1. In the **Variables permitted in table** box, select the variable that you want to add to the table. The variables in this list were assigned to the **Table variables** role.
2. Drag the selected variable to the **Preview** box. The values of the first variable that you drag comprise the variables of the crosstabulation table. The values of the second variable that you drag comprise the rows of the crosstabulation table.

   If you drag more than two variables to the **Preview** box, multiple crosstabulation tables are generated, one for each combination of values for the third (and later) variables that you drag.

   Note: After you add a variable to the table, it is removed from the **Variables permitted in table** box.

To remove a variable from the table, select the variable in the **Preview** box, and drag it to the **Variables permitted in table** box.

To start a new table, select **< select to begin defining a new table >** in the list of defined tables.

Table Analysis: Selecting Cell Statistics

In the selection pane, click **Cell Statistics** to access these options.

The selected statistics are shown in each cell of the crosstabulation table.

Option Name	Description
Cumulative column percentages	displays the cumulative column percentage in each cell.
Row percentages	displays the row percentage for each cell.
Column percentages	displays the column percentage for each cell. By default, these statistics are included in the output.
Cell frequencies	displays the frequency count for each cell. By default, these statistics are included in the output.
Cell percentages	displays the percentage, row percentage, and column percentage in crosstabulation tables, or percentages and cumulative percentages in one-way frequency tables.
Missing value frequencies	displays the frequencies for the missing values.
Cell contribution to Pearson chi-square	displays each cell's contribution to the total Pearson chi-square statistic.
Cell frequency deviation from expected	displays the deviation of the cell frequency from the expected value for each cell.
Expected cell frequency	displays the expected cell frequency for each cell.
Percentage of total frequency	displays the percentage of total frequency on *n*-way tables when *n*>2.
Include percentages in the data set	displays the percentages for the columns and rows in the output data set.

Table Analysis: Specifying Table Statistics

Performing Association Tests

In the selection pane under the Table Statistics heading, click **Association** to access these options.

Note: If you select **Exact p-values**, the exact *p*-value calculations can consume large amounts of memory and processing time; use the **Limit computation time** option to specify a time limit for each exact *p*-value computation.

Option Name	Description
Tests of association	You can specify these options: ■ **Chi-square tests** calculates the chi-square tests of homogeneity or independence and measures of association based on chi-square. The tests include the Pearson chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-square. For 2×2 tables, this test includes Fisher's exact test and the continuity-adjusted chi-square. ■ **Exact p-values** computes the exact *p*-values for the following statistics: 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. ■ **Fisher's exact test for r x c tables** computes Fisher's exact test for tables that are larger than 2×2. This test is also known as the Freeman-Halton test.
Measures of association	You can specify these options: ■ Measures computes several measures of association and their asymptotic standard errors (ASE). The measures include gamma, Kendall's tau-*b*, Stuart's tau-*c*, Somers' *D (C|R)*, Somers' *D (R|C)*, the Pearson and Spearman correlation coefficients, lambda (symmetric and asymmetric), and uncertainty coefficients (symmetric and asymmetric). For 2×2 tables, this measure also provides the odds ratio, column 1 relative risk, column 2 relative risk, and corresponding confidence limits. ■ **Exact p-values and confidence bounds for odds ratio** computes the exact *p*-values for the Pearson correlation and the Spearman correlation, and the odds ratio confidence limits for 2×2 tables. ■ **Test that the measure equals zero** computes asymptotic tests for gamma, Kendall's tau-*b*, Stuart's tau-*c*, Somers' *D (C|R)*, Somers' *D (R|C)*, the Pearson correlation, and the Spearman correlation. ■ **Risk differences for 2 x 2 tables** provides estimates of risks (or binomial proportions) and risk differences for 2×2 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. ■ **Relative risk for 2 x 2 tables** computes the relative risk measures and their confidence limits for 2×2 tables. These measures include the odds ratio and the column 1 and 2 relative risks.
Cochran-Mantel-Haenszel statistics	To compute all of the Cochran-Mantel-Haenszel statistics, select the **CMH statistics** check box. These statistics include the CMH correlation statistic, the row mean scores (ANOVA), and the adjusted relative risks and odds ratios. The default score type is **`Table`**. To change the score type, use the **Score type** option.

Performing Agreement Tests

In the selection pane under the Table Statistics heading, click **Agreement** to access these options.

You can select these tests and measures of agreement for *n*x*n* tables.

Option Name	Description
Measures	computes tests and measures of classification agreement for square tables. This option provides McNemar's test for 2×2 tables and Bowker's test of symmetry for tables with more than two response categories. It also produces the simple kappa coefficient, the weighted kappa coefficient, the asymptotic standard errors for the simple and weighted kappas, and the corresponding confidence limits. When there are multiple strata and two response categories, this option also computes Cochran's Q test.
Exact p-values	computes the exact *p*-values for McNemar's test for 2×2 tables, the simple kappa coefficient, and the weighted kappa coefficient. **Note:** If you select **Exact p-values**, the exact *p*-value calculations can consume large amounts of memory and processing time; use the **Limit computation time** option to specify a time limit for each exact *p*-value computation.
Test that the measure equals zero	computes asymptotic tests for the simple kappa coefficient and the weighted kappa coefficient.

Performing Ordered Differences Tests

In the selection pane under the Table Statistics heading, click **Ordered Differences** to access these options.

Option Name	Description
Jonckheere-Terpstra test	computes 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.
Exact p-values	computes the exact *p*-value for the Jonckheere-Terpstra test. **Note:** If you select **Exact p-values**, the exact *p*-value calculations can consume large amounts of memory and processing time; use the **Limit computation time** option to specify a time limit for each exact *p*-value computation.

Performing Trend Tests

In the selection pane under the Table Statistics heading, click **Trend Test** to access these options.

Option Name	Description
Cochran-Armitage test	computes 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 contingency 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.
Exact p-values	computes the exact *p*-value for the Cochran-Armitage test for trend. **Note:** If you select **Exact p-values**, the exact *p*-value calculations can consume large amounts of memory and processing time; use the **Limit computation time** option to specify a time limit for each exact *p*-value computation.

Setting Computation Options

In the selection pane under the Table Statistics heading, click **Computation Options** to access these options.

Option Name	Description
Score type	If you are performing association, agreement, or trend tests, select a score type. The score type specifies the type of row and column scores to use with the Mantel-Haenszel chi-square statistic, Cochran-Mantel-Haenszel statistics, Pearson correlation, Cochran-Armitage test for trend, and weighted kappa coefficient. You can choose from these options: ■ **Table** - For numeric variables, table scores are the values of the row and column levels. If the row or column variables are 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** - Rank scores are defined by row scores and column scores. ■ **RIDIT** - Ridit scores are derived from rank scores. ■ **Modified RIDIT** - Modified ridit (MODRIDIT) scores represent the expected values of the order statistics for the uniform distribution on (0,1). Modified ridit scores are derived from rank scores.
Exact computations	If you are calculating exact *p*-values in any of your tests, calculating exact *p*-values can consume large amounts of memory and processing time. In the **Exact computations** area, select the **Limit computation time** check box and specify a time limit (in seconds) for the computation of each *p*-value for each table. The default is `900` seconds. To compute the Monte Carlo estimates of the exact *p*-values instead of directly computing the exact *p*-values, select the **Use Monte Carlo estimation** check box. 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.
Include missing values in calculations	You can choose to include missing values in the calculations of percentages and other statistics.
Confidence level	You can specify a confidence level for the calculations. By default, the confidence level is a 95% confidence limit.
Order values by	You can specify the order in which the values of the variables in the crosstabulation tables are reported. **Order values by** can be set to the following options: ■ **Data set order** to order the values by their appearance in the data set ■ **Formatted values** to order the values by their ascending formatted values ■ **Descending frequencies** to order the values so that the levels with the highest frequency counts come first ■ **Unformatted values** to order the values by their unformatted values. This is the default.

Table Analysis: Setting Results Options

Saving Cell Statistics

In the selection pane under the Results heading, click **Cell Stat Results** to access these options.

Option Name	Description
Saving the output data	Select the check box for the desired crosstabulation table or tables. The table that is listed corresponds to the variable that is assigned to the **Table variables** role. In the **Cell statistics data set** box, specify the data set in which to save the cell statistics for the selected table. The name of the data set changes for each table that you select. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Suppress all displayed output	Does not display the results.

Saving Table Statistics

In the selection pane under the Results heading, click **Table Stat Results** to access these options.

Option Name	Description
Save the statistics for the following table	saves the table statistics for one crosstabulation table. In the **Table statistics data set** box, specify the data set in which to save the table statistics for the selected table. The name of the data set changes for each table that you select. Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list. The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.
Display tables in a list format	displays the entire multiway table in one table, instead of displaying a separate two-way table for each stratus. Note: The **Display tables in a list format** option is ignored if you select any option that generates a table statistic.
Suppress all displayed output	does not display the results.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

90

Tile Chart

About the Tile Chart Task

The Tile Chart task enables you to create a tile chart so that you can view a large quantity of hierarchical data in a limited space. Each unique category combination is represented by a rectangular tile whose size and color are determined by response variables. These tiles are placed in a hierarchical arrangement.

A tile chart is a rectangular tree map. Because the tile chart is displayed as a tree, it is especially useful in an interactive environment, such as the web. Here are the advantages to tile charts in an interactive environment:

- To see small sections of the original chart in more detail, you can use the drill-down (or zooming) functionality.
- When you position your mouse pointer over a tile, data tips show the metrics for a particular tile.
- You can link to a report or web page from the tiles on a particular level. When you click a tile in that level, the report or web page that is associated with that tile opens.

Only the ActiveX, Java, ActiveX Image, and Java Image graph formats are valid with the Tile Chart task. If the current graph format is not one of these formats, the Tile Chart task uses the ActiveX graph format.

For HTML results, some functionality is available only for specific graph formats.

- The drill-down functionality is available if you selected ActiveX or Java as the format for the graphical output.
- Linking to a web page from your tile chart is available if you selected Java as the format for the graphical output.

You can specify the graph format and the format for your results in the Options dialog box.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	GTILE

Requirement Name	Procedure and Product Names
Required SAS products	Base SAS, SAS/GRAPH
Recommended additional SAS products	None

Tile Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Task Roles** to access these options.

After you have selected a data source, you can assign variables to roles.

Note: The available roles depend on the chart type that you selected.

Role Name	Description
Classification variables	specifies the hierarchical category (or ID) variables. A classification variable is required, and you assign a maximum of six variables to this role. If you assign multiple variables to this role, then the order of the variables defines the hierarchy. For example, if the order of the variables in the Classification variables role is Country, Region, and Division, then the resulting hierarchy is Country > Region > Division.
Color analysis variable	specifies the variable values that determine the colors of the tiles. The smallest color response is assigned the minimum color list value, and the largest color response is assigned the maximum color list value. Other color responses are assigned linearly in a gradient between colors in the list. You can assign only one variable to this role. **Note:** If no variable is assigned as the color analysis variable, then the variable that you assign to the **Size analysis variable** role is also used as the color analysis variable.
Size analysis variable	specifies the variable values that determine the relative size of the tiles. The total size of all tiles covers the entire tile area. Missing, zero, and negative values cause their tiles to be ignored. A size analysis variable is required, and you can assign only one variable to this role.
Group charts by	The values of the column that you assign to this role determine the number of graphs that are created. A separate graph is created for each unique value. You can assign more than one variable to this role.

Tile Chart: Setting Appearance Options

Specifying the Appearance of the Tile Chart

In the selection pane under the Appearance heading, click **Tile Chart** to access these options.

Role Name	Description
Tile layout	Specify the tile layout in the chart. You can choose from these options: ■ **Flow** divides the tile layout over several, aligned rows. In the chart, you read the ordered tiles from left to right and then from top to bottom. ■ **Square** is the classic rectangular tree map. Usually, the tiles are arranged from largest to smallest from the lower left corner of the chart to the upper right corner of the chart. Sorting on the data variables is ignored, and the shape of the tiles is as square as possible. ■ **Toggle** splits the direction for each level of the hierarchy that resulted from the variables that you assigned to the **Classification variables** role. The highest level of the hierarchy is split into vertical slices of the available space and is displayed in sorted order along a single axis. At lower levels, the split direction is changed (or toggled), so that the second level is split horizontally.
Specify classification level	specify the number of levels to display in the results. For example, if the order of the variables in the **Classification variables** role is Country, Region, and Division, then the resulting hierarchy is Country > Region > Division. The number of levels that are displayed in the tile chart depends on the value of the **Specify classification level** option. ■ If this option is set to 1, then one level of the hierarchy is displayed in the results. The tile chart displays the tiles for each distinct country (such as Canada, Germany, and the United States). If you drill down on the Canada tile, then the tile chart displays the tiles for the two regions (West and East) in Canada. ■ If this option is set to 2, then two levels of the hierarchy are displayed in the results. The tile chart displays the tiles for each country and region (such as Canada West, Canada East, Germany West, Germany East, United States West, and United States East). If you drill down on the Canada tile, then you see tiles for all the divisions in the Canada West and Canada East regions. By default, all the levels of the hierarchy are displayed.
Specify colors	customizes the color ramp for the tile chart. A color ramp is the distribution of colors across a range of data values. By default, you must specify at least two colors. These colors are the start and end colors for the color ramp. By default, the colors in the color ramp are evenly distributed across the gradient legend. You can use color points to specify how the colors should be distributed across the data. To specify color points, click **Edit**. In the Edit Color Points dialog box, values can range from 0 to 1 and must be in ascending order. For example, if you specified three colors for the color ramp, then you might specify 0, .75, and 1 as color points. You can specify the same value for different color points. For example, if you specified five colors for the color ramp, then you might specify 0, .2, .4, .4, and .6 as color points. You must specify the same number of color points as the number of colors in the color ramp.

Setting Legend Options

To specify the start and end points for the legend, select the **Specify legend range** check box. Suppose the values in your data range from 0 to 100,000. However, on the tile chart, you want the color mapping to highlight the 22,000 to 26,000 values. In this example, you would specify 22,000 as the minimum (or start) value and 26,000 as the maximum (or end) value. The tiles with values less than 22,000 or greater than 26,000 are gray. The tiles with values of 22,000 to 26,000 use the color ramp that you specified for the tile chart.

By default, the color mapping in the legend uses the colors in the response data. However, you can specify whether the color mapping should be symmetrical on either side of a fixed value by selecting the **Set legend color mapping to be symmetrical about a fixed value** check box. This option enables you to specify the

midpoint in the color scale. By default, the point of symmetry or midpoint is 0. You can specify a new point of symmetry or midpoint.

Tile Chart: Setting the Advanced Options

In the selection pane, click **Advanced** to access these options.

Note: You can link to web pages from your tile chart only if you have selected Java as the format for your graphical output and HTML as the format for your results in SAS Enterprise Guide. To specify the graph format and the format for your results, see the**Results** ⇨ **Graph** panel in the Options dialog box. Because of the Java requirement, links are not supported in the SAS Add-In for Microsoft Office.

You can add links to web pages from your tile chart. A link is available from the pop-up menu that appears when you right-click a tile in a level for the selected link variable. (The link variables are the variables that you assigned to the **Classification variables** role.) The level where these links appear depends on whether you specified a classification level.

- If you do not specify a classification level, then links are available from all levels of the tile chart. For example, if the hierarchy for the tile chart is Country > Region > Division, then the links are available when you right-click the tiles in the Country, Region, and Division levels.
- If you specify a classification level, then links are available only in the lowest levels of the tile chart. For example, the hierarchy for the tile chart has three levels: Country, Region, and Division. If the classification level is 2, then the links are available in the two lowest levels, Region and Division. If the classification level is 1, then the links are available only from the lowest level, Division.

In the Tile Chart task, the link variables are listed in the **Available link variables** box. The task automatically assigns a substitution number to each link variable. This number appears next to its associated variable in the **Available link variables** box. You can include this substitution number in the URL for the web link or label to create a unique link for each value of a link variable.

For example, in your data, the values of Country are fr (France), jp (Japan), and uk (the United Kingdom). If 1 is the substitution number that is automatically assigned for the Country variable and you specify http://www.sales.{&1}.com as the URL for the web link, these links are available from the tile chart:

- http://www.sales.fr.com
- http://www.sales.jp.com
- http://www.sales.uk.com

To include links in your chart:

1. Specify the label for the link. To create a unique label for each value of the link variable, you can include the substitution number of the link variable for the label. To add the substitution number for a link variable in the label, click `Insert Variable`.

   Here are examples of how to specify the link label:

   - `State Web page for {&1}`
   - `Wikipedia Web page for {&2}`

2. Specify the URL for the web page. You can create a unique web link for each value of the link variable by including the substitution number for the link variable in the URL. The syntax for specifying the link variable is {&*substitution-number*}. This substitution value can appear anywhere in the URL for the web link. To add the substitution number for a link variable in the URL, click **Insert Variable**.

   Here are examples of how to specify the URL for the web link:

   - http://www.state.{&1}.us

- http://en.wikipedia.org/wiki/{&2}

3 Click **Add** to add this Web link to the list of Web links for the current tile chart.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1 Select the title or footnote in the **Section** box.

2 Clear the **Use default text** check box.

3 Edit the text of the title or footnote in the text box.

Note: You can use macro variables in titles and footnotes.

91

Transpose

About the Transpose Task

The Transpose task turns selected columns of an input data set into the rows of an output data set. If you do not use grouping variables, then each selected column is turned into a single row. If you use grouping variables, then the selected columns are divided into subcolumns based on the values of the grouping variables, and each subcolumn is turned into a row of the output data set.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	TRANSPOSE
Required SAS products	Base SAS
Recommended additional SAS products	SAS/STAT

Transpose: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Transpose variables	Each column that you assign to this role becomes one or more rows of the output data set. If you do not select any grouping variables, then an entire column is turned into a single row. If you select one or more grouping variables, then the grouping variables are used to segment each column into subcolumns, each of which is turned into a row. In this case, a column is transposed to the number of rows that is equal to the number of groups that are defined by the grouping variables. You must assign at least one column to the **Transpose variables** role. To select a grouping variable, assign a column to the **Group analysis by** role.
Copy variables	Each column that you assign to this role is copied directly from the input data set to the output data set without being transposed. Because these columns are copied directly to the output data set, the number of rows in the output data set is equal to the number of rows in the input data set. The output data set is padded with missing values if the number of rows in the input data set does not equal the number of variables that it transposes.
New column names	If you assign an input column to this role, then the values in that column are used to form the suffixes of the data column headings in the output data set. If you do not assign an input column to this role, then integers are used to form the suffixes of the data column headings in the output data set. To specify the prefix for the data column headings, click **Options** in the selection pane, and select the **Use prefix** check box. Then enter the prefix in the **Column name prefix** box. In general, a column that is used to provide column names for the output data set should not contain any duplicate values, since this would cause duplicate column names. If you want to use a column that contains duplicate values, select the column after you assign it, then select the **Allow duplicates** check box that appears to the right. In this case, only the last row that contains a given value will be used in the transposition. If you use a column that contains duplicate values and you do not select the **Allow duplicates** check box, then the transposition cannot be carried out.
Group analysis by	Each variable that you assign to this role is used to segment the about-to-be-transposed columns into subcolumns that will be transposed separately. Each subcolumn, defined by a set of values of the grouping variables, becomes a row of the output data set.
New column labels	This role becomes available if you assign an input column to the **New column names** role. If you assign an input column to this role, then the values in that column are used as the labels of the data variables in the output data set. If you do not assign an input column to this role, then the labels of the data variables in the output data set will be identical to the names of those variables.

Transpose: Setting Analysis Options

In the selection pane, click **Options** to access these options.

Each row of the output data set includes the name and label of the variable in the input data set. You can specify a heading for the output column that contains these variable names or variable labels. The name or label can include special characters, leading numbers, and white space, but it cannot exceed 32 characters. The default heading for the variable name column is **Source**. The default heading for the variable label column is **Label**.

The headings of all the data columns of the output data set begin with the same prefix. The prefix can include special characters, leading numbers, and white space, but it cannot exceed 32 characters. The default prefix is **Column**.

Note: Specifying a prefix is optional if you assign a variable to the New column names role. If you do not assign a variable to this role, a prefix is required, and the **Use prefix** check box is not available.

Transpose: Setting Results Options

In the selection pane, click **Results** to access these options.

You can designate a different name for the output data set.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

92

u Chart

About the u Chart

The Shewhart control chart is a graphical and analytical tool for determining whether a process is in statistical control. The u Chart task creates u charts for the numbers of nonconformities (defects) per inspection unit in the subgroup samples that contain arbitrary numbers of units.

You might want to use this task to determine the number of defects per inspection unit that resulted from a manufacturing process. For example, suppose that a textile company uses a u chart to monitor the number of defects per square meter of fabric. The fabric is spooled onto rolls as it is inspected for defects. Each piece of fabric is one meter wide and 30 meters long. The defect counts for 20 rolls are saved in a SAS data set, which is used to create the u chart. Each point on the u chart represents the number of nonconformities per inspection unit for a particular subgroup. For example, the value that is plotted for the first subgroup is 12/30=0.4 (because there are 12 defects on the first roll and the roll contains 30 square meters of fabric). If none of the points exceed the control limit (which is 3 sigma by default), the u chart indicates that the fabric manufacturing process is in statistical control.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	SHEWHART
Required SAS products	Base SAS, SAS/QC, SAS/GRAPH
Recommended additional SAS products	SAS/STAT

u Chart: Assigning Variables to Analysis Roles

In the selection pane, click **Data** to access these options.

By default, the data source that you selected before opening the task is the input data source for the task. The name of the input data source appears in the **Data source** field. Any filters that were specified in the data source appear in the **Filter** field. To change or filter the input data source, click **Edit**.

After you have selected a data source, you can assign variables to roles.

Role Name	Description
Number of nonconformities	specifies the variables that contain the number of nonconformities. You must assign at least one variable to this role
Subgroup identifier	specifies the variable that identifies the subgroups in the data. The values of the variable indicate how the observations in the input table are arranged into rational subgroups. Typically, here are the values: ■ the indices that give the order in which the subgroup samples were collected ■ the dates or times at which the subgroup samples were collected ■ the labels that uniquely identify the subgroup samples A subgroup identifier is required, and you can assign only one variable to this role. To sort the data by the identified subgroups, select the **Sort by subgroup** check box.
Subgroup sample size	specifies the subgroup sample sizes as the values of a variable. The subgroup sample size is required, and you assign only one numeric variable to this role. Select the **Use numeric value for subgroup (instead of a variable)** check box when the subgroup sample size is fixed and you need to specify a constant value for the size. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** This role is not required if you select the **Use numeric value for subgroup (instead of a variable)** check box on the **Options** pane.
Group analysis by	produces separate analyses on rows in groups that are defined by the grouping variables. By default, the data is sorted in ascending order by the values of these grouping variables. **Note:** To prevent the input data set from being sorted, clear the **Sort by variables** check box. **Note:** You cannot group analyses by a variable that you have already selected for an analysis role.
Block variables	specifies up to two variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

Control Charts: Specifying Control Limits

In the selection pane, click **Control Limits** to access these options.

Option Name	Description
Sigma limits	specifies the width of the control limits as a multiple of the standard error (sigma) of the summary subgroup statistic that is plotted on the chart. The width must be positive. The default multiple is `3`.
Select computation method	specifies the computation method for the control limits. You can choose from these options: ■ **Compute the control limits from the active data** ■ **Compute the control limits from the selected data set** To specify the data set that contains the control limits or the parameters from which the control limits can be computed, click **Browse**. Select the data set that you want to use and click **Open**. **Note:** The data set that you select must be a LIMITS= data set. For more information about the variables that are required in a LIMITS= data set, see the procedure documentation for the chart that you are creating. To plot the summary statistics for all subgroups, select the **Plots summary statistics for all subgroups, regardless of whether the subgroup sample size equals the nominal control limit sample size** check box. You might want to plot these statistics when almost all of the subgroups have a common sample size, and when you want to display the fixed (rather than varying) control limits that correspond to the nominal sample size *n*. If you select this option and your sample sizes for your subgroups differ significantly, then the interpretation of the control limits is meaningful only for those subgroups whose sample size equals *n*. To add special markers for the data points where the sample size does not equal `n`, select the **Add special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits** check box. ■ **Specify the control limits** In the **Upper**, **Control**, and **Lower** fields, specify the upper control limit, the central line, and the lower control limit, respectively. If your subgroup sample sizes vary, then you must specify a nominal sample size. **Note:** If you assign any variables to the **Group analysis by** role, then you cannot enter control limit values.

Control Chart: Selecting Tests to Perform

In the selection pane, click **Tests** to access these options.

Note: These tests do not apply to the control limits that are not three sigma limits or that vary with the subgroup sample size.

Option Name	Description
Select test	Use the check boxes to request one or more tests for special causes, which are also known as runs tests, pattern tests, and Western Electric rules. These tests detect particular nonrandom patterns in the points that are plotted on the primary control chart. The occurrence of a nonrandom pattern is referred to as a signal. The tests can indicate the presence of a special cause of variation. When you select each test, a description of the test appears in the **Description** box.
Label	Use the text boxes to enter the labels for points at which the tests for special causes are positive. The labels cannot exceed 16 characters. Each label appears at all the points in the test where a signal occurs.
Text identifying test signals	■ Use the drop-down list to select a color for the labels that you specified. ■ Use the **Display zone lines** check box to include lines that delineate zones A, B, and C in a primary chart. These zones are constructed by dividing the interval between the control limits into six equally spaced subintervals. ■ Use the **Override 3 sigma limit** check box to test for special causes when you specify a sigma control limit other than the default multiple of 3. To change the sigma limit, click **Control Limits** in the selection pane. ■ Use the **Apply test to overlapping patterns of points** check box to apply tests for special causes to the overlapping patterns of points.

Control Charts: Setting Plot Appearance Options

Setting Axis Options

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Axes** to access the options.

You can set these options:

- the color of the axes
- the width of the axes
- the color of the tick marks
- a label for the horizontal and vertical axes

Adding Reference Lines

In the selection pane under the **Appearance** ⇨ **Axes** heading, click **Horizontal** or **Vertical** to access the options.

To use reference lines, select **Use reference lines**. You can format the reference line and specify the location for each reference line.

Setting Other Plot Options

In the selection pane under the Appearance heading, click **Options** to access these options.

Option Name	Description
Inside control limits, Control limits, Connecting line segments, Segments outside limits, Areas outside the limits, and Frame	specifies the color for these chart elements. ■ **Inside control limits** specifies the color inside the lower and upper control limits. ■ **Control limits** specifies the color for the control limits and the central line. It also specifies the color of the labels for these lines. ■ **Connecting line segments** specifies the color for the line segments that connect points on the chart. ■ **Segments outside limits** specifies the color for the plotting symbols and the portions of connecting line segments that lie outside the control limits. This option is useful for highlighting out-of-control points. ■ **Areas outside the limits** specifies the fill color for the areas outside the control limits that lie between the connected points and the control limits and are bounded by connecting lines. This option is useful for highlighting out-of-control points. ■ **Frame** specifies the colors for filling the rectangle that is enclosed by the axes and the frame.
Symbol	specifies the symbol for the data points.
Number of display pages	specifies the number of pages to use to display the chart.
Moving range	Specifies the number of consecutive measurements from which the moving ranges are computed. The specified value should be between 2 and 10 (inclusive). The default value is **2**. The range of 2 to 10 is specific to the task. If you are using the SHEWHART procedure in SAS, then you can specify a higher value. For more information, see the SHEWHART procedure in *Base SAS Procedures Guide*. **Note:** This option is available only for the Individual Measurements Chart task.
Suppress default subgroup sample size legend	suppresses the default legend for subgroup sample sizes. This option is useful when the subgroup sample sizes are constant and are equal to the control limit sample size, because the control limit sample size automatically appears in the upper-right corner of the chart.
Use numeric value for subgroup (instead of variable)	specifies a constant value for the size. Use this option when the subgropu sample size is fixed. Enter the value to use in the **Subgroup sample size** box. The subgroup sample size must be a value between the maximum value of the process variable and 99999 (inclusive). **Note:** If you assign a variable to the **Subgroup sample size** role, then this option is unchecked and the variable that you assigned to this role is used for the subgroup sample size. **Note:** This option is available only for the np, p, and u Chart tasks.

Option Name	Description
Omit secondary chart	prevent a secondary chart from being displayed. You typically use this option with an Individual Measurements chart to create a chart for individual measures and to suppress the accompanying chart for moving ranges. **Note:** This option is available only for the Individual Measurements Chart, the Mean and Range Chart, and the Mean and Standard Deviation Chart tasks.
Margin plot type	specifies the type of margin plot that will appear in the Individual Measurements chart. A margin plot is a univariate plot of the control chart statistics. You can select a digidot plot, a histogram, or a box-and-whisker plot. The plot appears either to the left or to the right of the Individual Measurements chart. The default selection is **No margin plot**. **Note:** This option is only available for the Individual Measurements Chart task.

Setting the Block Variable Options

In the selection pane under the Appearance heading, click **Block Options** to access these options.

Note: These options are available only if you have assigned a variable to the **Block variables** role.

Role Name	Description
Block label	specifies the labels for the block variables.
Legend position	specifies the vertical position of the legend for the values of the block variables.
Lengthy values	specify how to handle lengthy block variable values when there is insufficient space to display them in the block legend. By default, lengthy values are not displayed. If you select the **Reduce in height uniformly** option, the text size of the values is reduced in height so that they fit the legend. If you select the **Truncate on the right** option, lengthy values are truncated on the right until they fit the legend.
Label position	specifies the position of a block variable label in the block legend. The options that are available depend on the position of the legend. You can choose from these options: ■ **Above** places the label immediately above the legend. This is the default. ■ **Left** places the label to the left of the legend ■ **Right** places the label to the right of the legend. This option is not available if you have selected to position the legend above the control chart. You should specify **Left** or **Right** as the label position only with labels that are short enough to fit in the margins on each side of the chart. Labels that are too long are truncated.
Block label color	specifies the color for the text of the block label. **Note:** This option is available only if you select **Above** for the label position.
Display repeated values in legend	displays the block variable values for all subgroups. By default, only the first block variable in a block is displayed, and repeated block variable values are not displayed.

Control Charts: Saving the Analysis Results

In the selection pane, click **Tables** to access these options.

To save the analysis results, you can create output data sets that contain various statistics. Select the check boxes for the output data that you want to save permanently.

Role Name	Description
Subgroup statistics and control limit data	creates an output data set that contains the information that is plotted on the chart. This includes the values of the subgroup variable and their corresponding summary statistics and control limits. **Note:** This option creates an OUTTABLE= data set. You can read this data set back into SAS using the TABLE= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Subgroup statistics output data set	creates an output data set that contains the subgroup summary statistics. **Note:** This option creates an OUTHISTORY= data set. You can read this data set back into SAS using the HISTORY= option in the SHEWHART procedure. For more information, see the Help for the SHEWHART procedure.
Control limits output data set	creates an output data set that saves the control limits. You can use this data set when you select the computation method for the control limits in a subsequent use of this task.

Note: SAS Enterprise Guide searches the list of libraries that is defined in **Tasks** > **Output Library** area of the Options window and saves the output data in the first writable library from that list.

The SAS Add-In for Microsoft Office stores the output data in your Work library. To designate a different storage location for your output data, click **Browse**.

Specifying Titles and Footnotes

In the selection pane, click **Titles** to access these options.

To accept the default title and footnote text for the results, you do not need to do anything.

To specify your own titles and footnotes:

1. Select the title or footnote in the **Section** box.
2. Clear the **Use default text** check box.
3. Edit the text of the title or footnote in the text box.

   Note: You can use macro variables in titles and footnotes.

Viewing Properties

In the selection pane, click **Properties** to access these options.

You can view a summary of the properties for a task from the **Properties** panel. For some tasks, you can edit these properties by clicking **Edit**. The Properties dialog box appears.

From the **Properties** panel, you can view this information:

- the label for the task
- the server used to run the task
- the date and time that the task was created
- the last date and time that the task was modified
- the last time the task was run
- any limits on the data
- any prompts that were used
- the format of the results
- the format for the graphical output in the results

Note: If you change the label for an existing task, the new label is saved when you click **OK** in the Properties dialog box. The new label is used even if you cancel the task.

93

Update Library Metadata

About the Update Library Metadata Wizard

The Update Library Metadata wizard enables you to run a report on table definitions in the metadata repository or to update these table definitions. You can create a report that lists the library's contents based on the table definitions in the metadata repository or a report that lists the differences between the tables in the physical library and the table definitions in the metadata repository. In addition to running reports, you can also use the Update Library Metadata wizard to update, add, or delete table definitions in the metadata repository.

By default, SAS Enterprise Guide shows only those tables that are registered in the metadata. The Update Library Metadata wizard uses the METALIB procedure to programmatically register these tables in the metadata repository and to synchronize metadata for these tables.

Note: This task is available only in SAS Enterprise Guide.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	METALIB
Required SAS products	Base SAS
Recommended additional SAS products	None

Update Library Metadata: Select the SAS Server and Library

1. Select the server from the list of available servers.
2. Select a library.
3. Click **Next** to run a report or update the metadata for the selected library.

Update Library Metadata: Run a Report on the Metadata or Update the Metadata

1 Before you can run a report or update the metadata, you must select the SAS server and library.

2 Select one of these options:

Option Name	Description
Report on the contents of the library, as registered in the metadata repository	generates a report of the library's contents based on the table definitions in the metadata repository.
Report on the differences between physical tables and the metadata repository	generates a report that contains this information: ■ the table definitions in the metadata repository that would be updated to match the corresponding physical tables ■ the tables in the physical library that do not have table definitions in the metadata repository ■ the tables that no longer exist in the physical library but have table definitions in the metadata repository
Update and add table definitions in metadata repository with the actual tables and columns	performs these steps: ■ For tables in the physical library that do have metadata, the table definitions in the metadata repository are updated to match the corresponding physical tables. These updates include information about the table's columns, indexes, unique keys, foreign keys, and key associations. ■ For tables in the physical library that do not have metadata, those table definitions are added to the metadata repository.
Update only the existing table definitions in metadata with current column information	updates the table definitions in the metadata repository for those tables in the physical library that have metadata.
Delete obsolete entries from the metadata library (tables no longer exist)	deletes from the metadata repository the table definitions for tables that no longer exist in the physical library.

3 Before you can update or delete metadata, you might need to specify the user ID and password for an account that has administrative privileges. If no user ID and password are required, then these options are not available.

4 Click **Finish** to run the report or update the metadata in the library.

94

Upload to CAS Task

About the Upload to CAS Task

The Upload to CAS task enables you to load SAS 9.4 data to a Cloud Analytics Services (CAS) server. This task uses SAS/CONNECT to move data to a remote server. Starting with SAS 9.4M5, there are more efficient ways to move data to a CAS server. For example, you can use the CAS LIBNAME engine or PROC CASUTIL. For more information, see *SAS Cloud Analytic Services: User's Guide*.

Note: To run this task in SAS Enterprise Guide or the SAS Add-In for Microsoft Office, you first must configure your environment. For more information, see *SAS Intelligence Platform: Desktop Application Administration Guide*.

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	Base SAS, SAS Viya
Recommended additional SAS products	None

Upload Data to a SAS CAS Analytic Server

1. Select the CAS server.
2. Select the library where to save the data.
3. Specify the member name of the table. By default, the member name is the same as the name of the data set. You cannot change the member name if you are using distributed loading.
4. Specify where to locate the credentials for the selected table. If you are connected to a SAS Metadata Server, you can use the credentials stored in the metadata. If you are uploading local data to the CAS server, you must enter the username and password for the CAS server.

95

Upload Data Files to Server Task

About the Upload Data Files to Server Task

The Upload Data Files to Server task enables you to select one or more SAS data files (*.sas7bdat) from your local Microsoft Windows file system and copy them to a SAS server.

The task generates a log that summarizes the data files that were copied, the byte size, timings for each file and for all of the files as a group, and any errors that occurred during the transfer. By default, all data files are added as output data in your SAS Enterprise Guide project so that they can be used in the project.

Upload Data Files to Server: Select the SAS Data Files to Upload

1 Click **Add** to select the data files that you want to upload.

2 Select the name of the file that you want to copy.

   Note: Files must be SAS7BDAT files.

3 Click **Open** to select the data files that you want to copy.

   Note: All files that are copied are given lowercase names on the target server.

4 Click **Next** to specify the upload options.

Upload Data Files to Server: Specify the Upload Options

1 Before you can specify the upload options, you must first select the data files that you want to upload.

2 From the **Select SAS server** drop-down list, select the server to which you want to copy the data files that you selected.

3 From the **Select destination library** drop-down list, select the SAS library that you want to use.

4 If you want to overwrite any existing files on the SAS server that have the same name as the files that you are copying, select the **Overwrite existing files with the same name** check box. By default, existing files are not overwritten

5 If you do not want to add the data files that you are copying to your current SAS Enterprise Guide project, clear the **Add data files to your current project** check box. By default, the copied data files are added to the current project.

6 Click **Finish** to copy the data files.

Note: If the target library is a Base SAS engine library with a single file path, then the file transfer will go directly to that library path. If the target library is not a Base SAS engine library or it is a concatenated library, the file transfer goes first to a staging area in Work. Then it is copied to the target library by using PROC SQL.

96

Upload to LASR Task

About the Upload to LASR Task

The SAS LASR Analytic Server is an in-memory analytics platform specifically designed to accelerate analytic computations. The server is available in a non-distributed configuration for deployments that do not have large data volumes. The distributed server uses a distributed computing environment and parallel processing for analyzing large data sets. The Upload to LASR task enables you to upload data to a LASR Analytic Server so that you can access the data in SAS Visual Analytics.

Note: To run this task, SAS Enterprise Guide or the SAS Add-In for Microsoft Office must be connected to the same metadata server as SAS Visual Analytics.

For information about how to set up a connection to a SAS LASR Analytic Server, see *SAS Intelligence Platform: Desktop Application Administration Guide*

To run this task, you must meet these requirements:

Requirement Name	Procedure and Product Names
SAS procedures used	None
Required SAS products	Base SAS, SAS LASR Analytic Server
Recommended additional SAS products	SAS Visual Analytics

Upload Data to a SAS LASR Analytic Server

1. Select the library where to save the data.
2. Specify the member name of the table. By default, the member name is the same as the name of the data set. You cannot change the member name if you are using distributed loading.
3. Specify the directory path for the uploaded data. By default, each library has a predefined location.

Appendix 1

Built-In Tasks

Built-in Tasks and Associated SAS Procedures and Licenses

The following table lists the built-in tasks for SAS Enterprise Guide and SAS Add-In for Microsoft Office. Each task lists the primary SAS procedures for that task and the SAS software licenses that are required to run a task.

Note: Your version of SAS Enterprise Guide or SAS Add-In for Microsoft Office might not include all of the tasks listed here. The availability of a task depends on whether the necessary SAS software is licensed and installed at your site. For more information, contact your system administrator.

If your site licenses and installs SAS Studio, additional tasks might be available. For more information, see SAS Studio tasks and associated SAS procedures and licenses.

Built-In Task	SAS Procedures Used	SAS License Required
Append Table	SQL	Base SAS
Area Plot	GPLOT	Base SAS, SAS/GRAPH
ARIMA Modeling and Forecasting	ARIMA	Base SAS, SAS/ETS
Assign Project Library	LIBNAME	Base SAS
Bar Chart	GCHART	Base SAS, SAS/GRAPH
Bar Chart Wizard on page 39	GCHART	Base SAS, SAS/GRAPH
Bar-Line Chart	GBARLINE	Base SAS, SAS/GRAPH
Basic Forecasting	FORECAST	Base SAS, SAS/ETS
Box Chart	SHEWHART	Base SAS, SAS/QC
Box Plot	GPLOT	Base SAS, SAS/GRAPH
Bubble Plot	GPLOT	Base SAS, SAS/GRAPH
c Chart on page 87	SHEWHART	Base SAS, SAS/QC

Built-In Task	SAS Procedures Used	SAS License Required
Canonical Correlation	CANCORR	Base SAS, SAS/STAT
CDF Plots	CAPABILITY	Base SAS, SAS/QC
Characterize Data	CONTENTS, UNIVARIATE, FREQ	Base SAS
Cluster Analysis	CLUSTER, FASTCLUS, TREE	Base SAS, SAS/STAT
Compare Data	COMPARE	Base SAS
Contour Plot	GCONTOUR, G3GRID	Base SAS, SAS/GRAPH
Correlations	CORR	Base SAS
Create Format	FORMAT	Base SAS
Create Format from Data Set on page 139	FORMAT	Base SAS
Create Map Feature Table on page 143	GPROJECT	Base SAS, SAS/GRAPH
Create Time Series Data	TIMESERIES	Base SAS, SAS/ETS
Data Set Attributes	DATASETS	Base SAS
Delete Data Sets and Formats	SQL, CATALOG	Base SAS
Discriminant Analysis	DISCRIM	Base SAS, SAS/STAT
Distribution Analysis	UNIVARIATE	Base SAS
Donut Chart	GCHART	Base SAS, SAS/GRAPH
Download Data Files to PC	SQL	Base SAS
Factor Analysis	FACTOR	Base SAS, SAS/STAT
Filter and Sort*	SQL	Base SAS
Forecast Studio Create Project	None	SAS Forecast Server
Forecast Studio Open Project	None	SAS Forecast Server
Forecast Studio Override Project on page 193	None	SAS Forecast Server
Generalized Linear Models	GENMOD	Base SAS, SAS/STAT
High-Performance Linear Regression	HPREG	Base SAS, SAS High Performance Analysis Server, SAS/STAT

Built-In Task	SAS Procedures Used	SAS License Required
High-Performance Logistic Regression	HPLOGISTIC	Base SAS, SAS High Performance Analysis Server, SAS/STAT
Histograms	CAPABILITY	Base SAS, SAS/QC
Import Data (Microsoft Excel, text files, Microsoft Access)*	DATA step, IMPORT	Base SAS, SAS/ACCESS to PC File Formats
Import JMP file*	DATA step, IMPORT	Base SAS, SAS/ACCESS to PC File Formats
Import SPSS file*	DATA step, IMPORT	Base SAS, SAS/ACCESS to PC File Formats
Import Stata file*	DATA step, IMPORT	Base SAS, SAS/ACCESS to PC File Formats
Individual Measurements Chart	SHEWHART	Base SAS, SAS/QC
JMP Stored Process Packager*	DATA step	JMP
Life Tables	LIFETEST	Base SAS, SAS/STAT
Line Plot	GPLOT	Base SAS, SAS/GRAPH
Line Plot Wizard	GPLOT	Base SAS, SAS/GRAPH
Linear Models	GLM	Base SAS, SAS/STAT
Linear Regression	REG	Base SAS, SAS/STAT
List Data	PRINT	Base SAS
List Report Wizard on page 281	REPORT	Base SAS
Logistic Regression	LOGISTIC	Base SAS, SAS/STAT
Map Chart	GMAP	Base SAS, SAS/GRAPH
Mean and Range Chart	SHEWHART	Base SAS, SAS/QC
Mean and Standard Deviation Chart	SHEWHART	Base SAS, SAS/QC
Mixed Models	MIXED	Base SAS, SAS/STAT
Model Scoring	None	SAS Enterprise Miner
Nonlinear Regression	NLIN	Base SAS, SAS/STAT
Nonparametric One-Way ANOVA	NPAR1WAY	Base SAS, SAS/STAT
np Chart	SHEWHART	Base SAS, SAS/QC

Built-In Task	SAS Procedures Used	SAS License Required
One-Way ANOVA	ANOVA	Base SAS, SAS/STAT
One-Way Frequencies	FREQ	Base SAS
SAS ODS Statistical Graphs	None	ODS Graphics Designer
p Chart	SHEWHART	Base SAS, SAS/QC
Pareto Chart	PARETO	Base SAS, SAS/QC
Pie Chart	GCHART	Base SAS, SAS/GRAPH
Pie Chart Wizard	GCHART	Base SAS, SAS/GRAPH
P-P Plots	CAPABILITY	Base SAS, SAS/QC
Prepare Time Series Data	EXPAND	Base SAS, SAS/ETS
Principal Components	PRINCOMP	Base SAS, SAS/STAT
Probability Plots	CAPABILITY	Base SAS, SAS/QC
Proportional Hazards	PHREG	Base SAS, SAS/STAT
Q-Q Plots	CAPABILITY	Base SAS, SAS/QC
Query Builder*	SQL	Base SAS
Radar Chart	GRADAR	Base SAS, SAS/GRAPH
Random Sample	SURVEYSELECT	Base SAS, SAS/STAT
Rank Data	RANK	Base SAS
Rapid Predictive Modeler	SAS Enterprise Miner macros	SAS Enterprise Miner
Recency, Frequency, and Monetary Analysis	SAS Macro	Base SAS
Regression Analysis of Panel Data	TSCSREG	Base SAS, SAS/ETS
Regression Analysis with Autoregressive Errors	AUTOREG	Base SAS, SAS/ETS
Scatter Plot	GPLOT, G3D	Base SAS, SAS/GRAPH
Scatter Plot Matrix	SGSCATTER	Base SAS, SAS/GRAPH
Show ODS Statistical Graph	SGDESIGN	Base SAS
Sort Data	SORT	Base SAS
Split Columns	TRANSPOSE	Base SAS

Built-In Task	SAS Procedures Used	SAS License Required
Stack Columns	TRANSPOSE	Base SAS
Standardize Data	STANDARD	Base SAS
Summary Statistics	MEANS	Base SAS
Summary Statistics Wizard	MEANS	Base SAS
Summary Tables on page 507	TRANSPOSE	Base SAS
Summary Tables Wizard	TRANSPOSE	Base SAS
Surface Plot	G3D	Base SAS, SAS/GRAPH
t Test	TTEST	Base SAS, SAS/STAT
Table Analysis	FREQ	Base SAS
Tile Chart	GTILE	Base SAS, SAS/GRAPH
Transpose	TRANSPOSE	Base SAS
u Chart	SHEWHART	Base SAS, SAS/QC
Update Library Metadata*	METALIB	Base SAS
Upload Data Files to Server*	SQL	Base SAS

Note: Tasks followed by an asterisk are available only in SAS Enterprise Guide.