



SAS Publishing

# The Analyst Application Second Edition



*The Power to Know.*

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. *The Analyst Application, Second Edition*. Cary, NC: SAS Institute Inc.

**The Analyst Application, Second Edition**

Copyright © 2003, SAS Institute Inc., Cary, NC, USA

ISBN 1-58025-991-X

All rights reserved. Produced in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**U.S. Government Restricted Rights Notice:** Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, June 2002

2nd printing, August 2003

Note that text corrections may have been made at each printing.

SAS Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at [support.sas.com/pubs](http://support.sas.com/pubs) or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

# Contents

---

Chapter 1. Overview . . . . .	1
Chapter 2. The Data Table . . . . .	25
Chapter 3. Managing Results in Projects . . . . .	67
Chapter 4. Customizing Your Session . . . . .	87
Chapter 5. Creating Graphs . . . . .	109
Chapter 6. Creating Reports . . . . .	151
Chapter 7. Descriptive Statistics . . . . .	181
Chapter 8. Hypothesis Tests . . . . .	209
Chapter 9. Table Analysis . . . . .	235
Chapter 10. Analysis of Variance . . . . .	267
Chapter 11. Regression . . . . .	299
Chapter 12. Sample Size and Power Calculations . . . . .	325
Chapter 13. Multivariate Techniques . . . . .	355
Chapter 14. Survival Analysis . . . . .	377
Chapter 15. Mixed Models . . . . .	393
Chapter 16. Repeated Measures . . . . .	413
Chapter 17. Details . . . . .	437
Appendix A. Summary of Tasks . . . . .	449
Subject Index . . . . .	473
Syntax Index . . . . .	481



# Acknowledgments

---

## Credits

---

### Documentation

Writing	Tim Arnold, Elizabeth Brownrigg, Nathan Curtis, Marty King, Maura Stokes, Lori Witham
Editing	Donna M. Sawyer, Maura Stokes
Editorial Support	Dee Doles
Production Support and Cover Design	Creative Solutions Division

---

### Software

Development	David DeNardis, Gregg Kemp, Julie LaBarr, Jeanne Martin, Katherine J. Roggenkamp, Jeff Sun, Wayne E. Watson
Testing	Daniel S. Adelsberg, Nathan A. Curtis, Wendy Hassler, Lori Witham, Ozkan Zengin

---

## Support Group

Quality Assurance

Mark F. Austin, Katherine Giacoletti, Kelly M. Graham

---

## About the Analyst Application

The Analyst Application is a point-and-click interface to basic statistical analyses in the SAS System. These analyses are performed primarily by using procedures in base SAS and SAS/STAT software, although some analyses are carried out with the use of specially written SAS macros.

---

## Documentation

This book describes the features of the Analyst Application and how to use it to perform typical analyses, but it is not intended to teach or describe the statistical methodology that is employed. You can find a description of the statistical techniques used in the *SAS/STAT User's Guide* and more tutorial-style information in several *Books By Users (BBU)* books. The pertinent books are listed in the back of each statistical task chapter.

---

## Software Requirements

The Analyst Application is available in Version 8 of the SAS System for the following platforms: Windows 95, Windows 98, Windows NT, UNIX workstations, OS/2, OpenVMS Alpha, and VMS VAX. Required are the following:

- Base SAS software, SAS/STAT, and SAS/GRAPH must be installed.
- SAS/ASSIST must be licensed.
- SAS/ACCESS must be installed in order to import external PC file formats.
- SAS/IML must be installed in order to produce confidence ellipses in the Correlations task and partial regression plots in the Linear Regression task; the selections are grayed out if SAS/IML is not available.





# Chapter 1

## Overview

### Chapter Contents

---

<b>Introduction</b> . . . . .	3
<b>Getting Started</b> . . . . .	4
Bring Up Analyst and Create a Sample Data Set . . . . .	4
Bring the Data into the Data Table . . . . .	5
Perform a Regression Analysis . . . . .	6
Your Results . . . . .	8
Saving a Project . . . . .	11
<b>Projects</b> . . . . .	12
Project Folders . . . . .	13
Nodes . . . . .	14
<b>Data Table</b> . . . . .	14
<b>Using the Mouse</b> . . . . .	15
Opening Nodes . . . . .	16
Selecting and Removing Variables . . . . .	16
<b>Accessing Tasks and Help</b> . . . . .	16
Menus . . . . .	16
Index . . . . .	17
Toolbar . . . . .	18
Getting Help . . . . .	19
<b>Graphs</b> . . . . .	20
<b>Reports</b> . . . . .	20

2 ♦ Chapter 1. Overview

Listing Reports . . . . .	20
Summary Reports . . . . .	21
<b>Statistical Tasks . . . . .</b>	<b>21</b>
Descriptive . . . . .	22
Table Analysis . . . . .	22
Hypothesis Tests . . . . .	22
ANOVA . . . . .	22
Regression . . . . .	23
Multivariate . . . . .	23
Survival . . . . .	23
Sample Size . . . . .	23

# Chapter 1

## Overview

---

### Introduction

The Analyst Application is a data analysis tool that provides easy access to basic statistical analyses. The application is intended for students and researchers as well as experienced SAS users and statisticians.

This interface takes a task-oriented approach to produce analyses and associated graphics. You can compute descriptive statistics, perform simple hypothesis tests, fit statistical models with regression and analysis of variance, and perform survival analysis as well as some multivariate analyses.

Most of the tasks provide access to analyses performed by SAS/STAT software, but some provide analyses not currently available with SAS/STAT procedures, such as certain hypothesis tests and basic sample size and power computations. In addition, you can produce many types of graphs, including histograms, box-and-whisker plots, probability plots, contour plots, and surface plots.

The Analyst Application enables you to input data in many ways, including opening data from external sources such as Excel files, inputting SAS data sets, or manually entering the data yourself. The data are displayed in a data table in which columns correspond to variables and rows correspond to observations or cases. You can edit individual elements in the data table, and you can create new columns and rows. You can also perform a number of other data manipulations such as subsetting the data, performing transforms, recoding, and stacking and splitting columns.

Once the data are ready, you specify tasks from pull-down menus, a customizable toolbar, or an index of commonly used task descriptions. The analysis results and plots are presented in separate windows and managed by a tree-list structure called a project tree. The underlying SAS code used to produce the results is available as a node in the project tree, and results can

also be displayed in HTML form and viewed with a web browser. You can save projects and then recall them for further work.

---

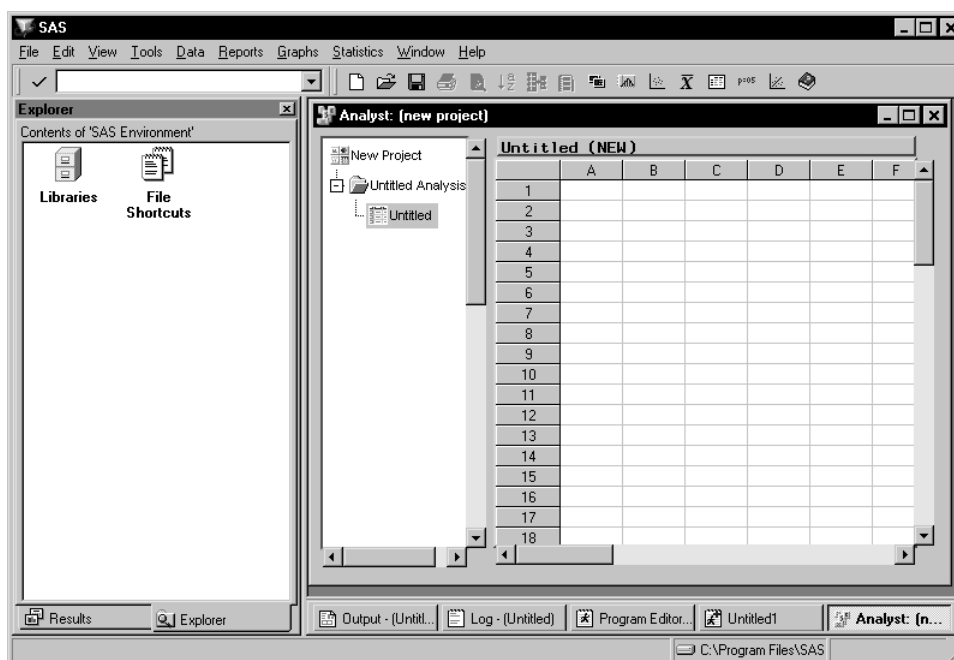
## Getting Started

In this example, you bring data into the Analyst data table, perform a regression analysis, and see the results in the project tree.

---

### Bring Up Analyst and Create a Sample Data Set

Select **Solutions** → **Analysis** → **Analyst** from the main SAS menu.



**Figure 1.1.** The Analyst Application with Explorer and Results Windows

You can close the Explorer and Results windows by clicking on the close box in the upper right corner of the windows. These windows are closed in the remaining examples in this book.

To use one of the sample data sets for this example, follow these steps:

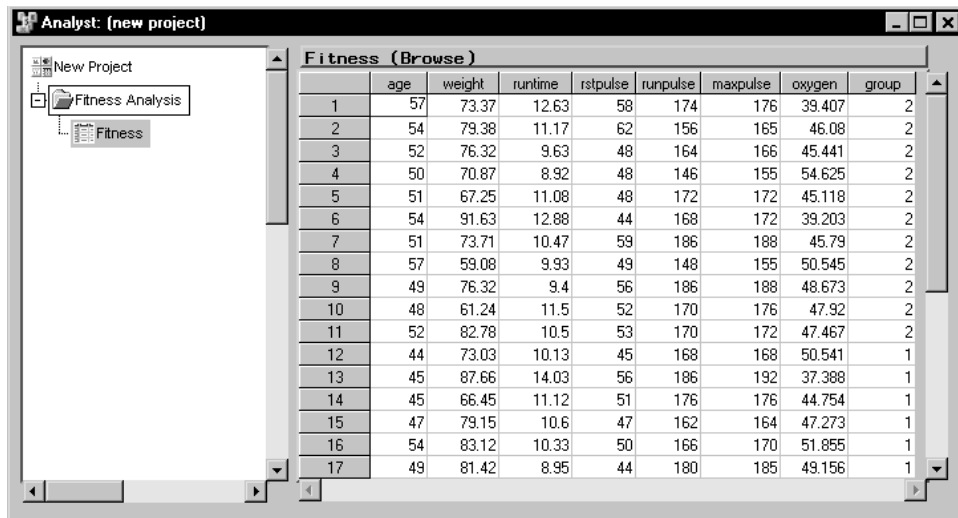
1. Select **Tools** → **Sample Data . . .**
2. Select **Fitness**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.

## Bring the Data into the Data Table

To bring the sample data into the data table, follow these steps:

1. Select **File** → **Open By SAS Name . . .**
2. Select **Sasuser** from the list of **Libraries**.
3. Select **Fitness** from the list of members.
4. Click **OK** to bring the sample data into the data table.

When you bring the **Fitness** sample data into the data table, the **Fitness Analysis** folder, with a **Fitness** table node representing the data table, appears in the project tree.



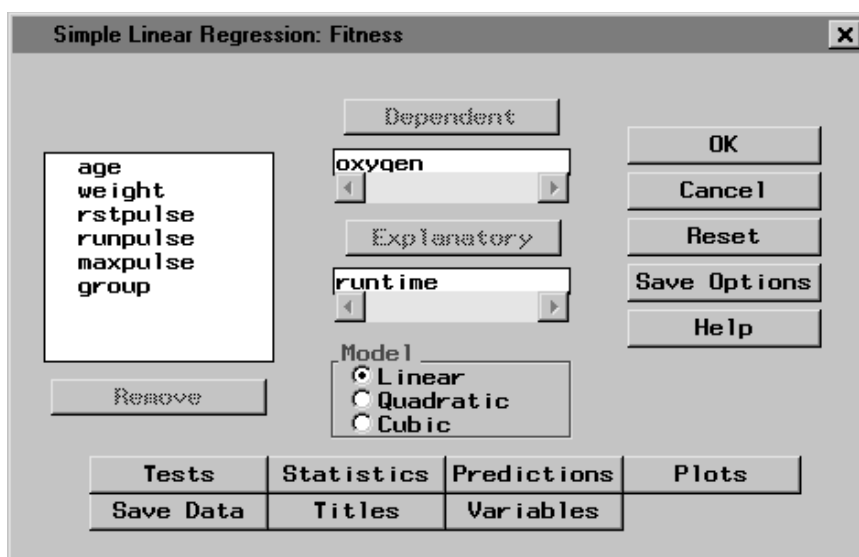
	age	weight	runtime	rstpulse	runpulse	maxpulse	oxygen	group
1	57	73.37	12.63	58	174	176	39.407	2
2	54	79.38	11.17	62	156	165	46.08	2
3	52	76.32	9.63	48	164	166	45.441	2
4	50	70.87	8.92	48	146	155	54.625	2
5	51	67.25	11.08	48	172	172	45.118	2
6	54	91.63	12.88	44	168	172	39.203	2
7	51	73.71	10.47	59	186	188	45.79	2
8	57	59.08	9.93	49	148	155	50.545	2
9	49	76.32	9.4	56	186	188	48.673	2
10	48	61.24	11.5	52	170	176	47.92	2
11	52	82.78	10.5	53	170	172	47.467	2
12	44	73.03	10.13	45	168	168	50.541	1
13	45	87.66	14.03	56	186	192	37.388	1
14	45	66.45	11.12	51	176	176	44.754	1
15	47	79.15	10.6	47	162	164	47.273	1
16	54	83.12	10.33	50	166	170	51.855	1
17	49	81.42	8.95	44	180	185	49.156	1

**Figure 1.2.** Fitness Analysis Folder in Project Tree

## Perform a Regression Analysis

To run a simple linear regression on the Fitness data, follow these steps.

1. Select **Statistics** → **Regression** → **Simple ...** to select the Simple Regression task.
2. In the Simple Linear Regression dialog, select **oxygen** from the list and click on the **Dependent** button to designate oxygen consumption as the dependent variable. Select **runtime** from the list and click on the **Explanatory** button to designate the amount of time to run 1.5 miles as the explanatory variable.



**Figure 1.3.** Dependent and Explanatory Variables

3. To create a scatter plot of your results, click on the **Plots** button. In the **Predicted** tab, select **Plot observed vs independent**.

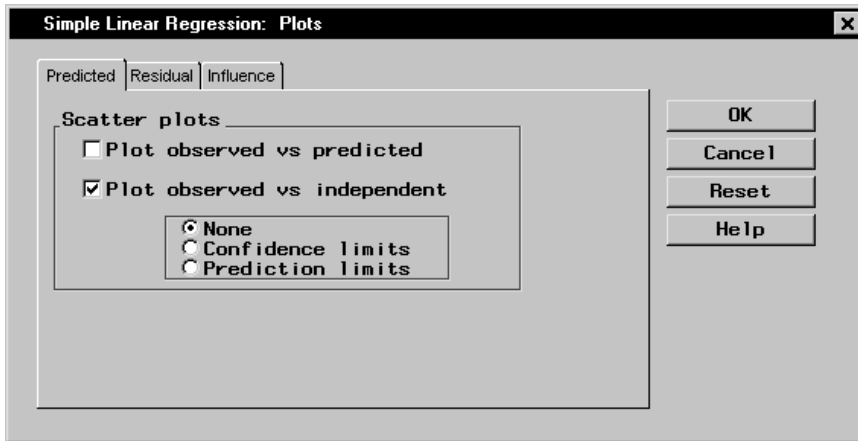


Figure 1.4. Scatter Plot

Click **OK**.

4. In the Simple Linear Regression dialog, click on the **Titles** button to specify a title for your results. In the first field of the Titles dialog, type **Speed vs. Oxygen Consumed**.

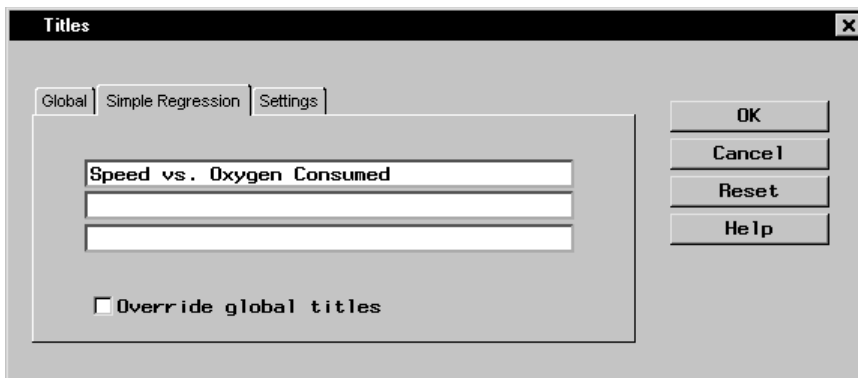


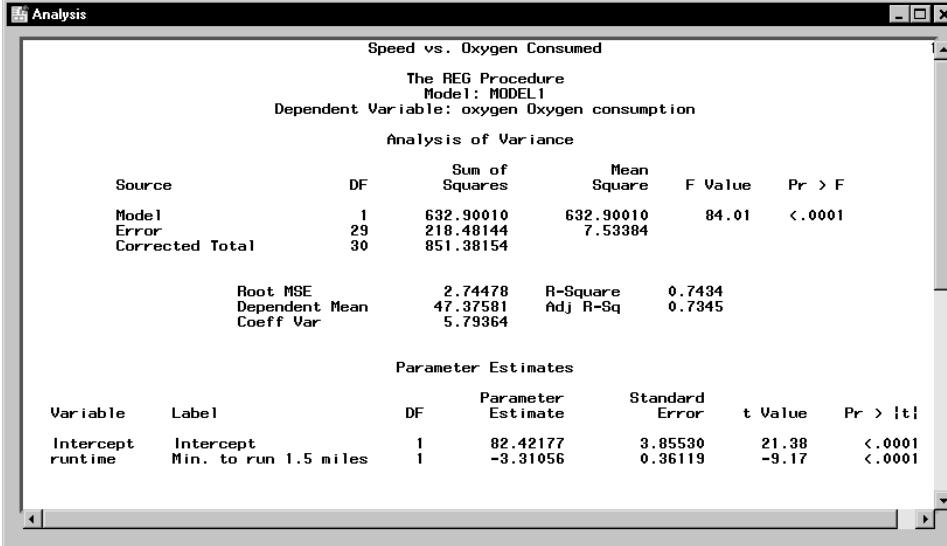
Figure 1.5. Title

Click **OK**.

- Click **OK** in the Simple Linear Regression dialog to generate the results.

## Your Results

When you click **OK** in the Simple Linear Regression dialog, the output from a simple linear regression is automatically displayed in the Analysis window. Drag the borders of the Analysis window until you can see all of the output.



Speed vs. Oxygen Consumed

The REG Procedure  
Model: MODEL1  
Dependent Variable: oxygen Oxygen consumption

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	632.90010	632.90010	84.01	<.0001
Error	29	218.48144	7.53384		
Corrected Total	30	851.38154			

Root MSE 2.74478 R-Square 0.7434  
Dependent Mean 47.37581 Adj R-Sq 0.7345  
Coeff Var 5.79364

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	82.42177	3.85530	21.38	<.0001
runtime	Min. to run 1.5 miles	1	-3.31056	0.36119	-9.17	<.0001

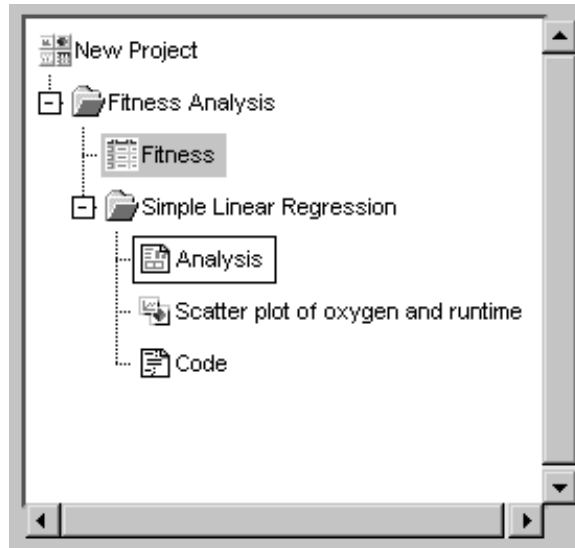
**Figure 1.6.** Simple Regression Output

This model might be considered minimally adequate with an R-square value of 0.7434; the negative coefficient for runtime indicates that the linear relationship between oxygen consumption and running time is a negative one.

You can save and print your results from this window. You can also copy your output to the Program Editor window, where you can copy it to the clipboard and paste it into other applications.

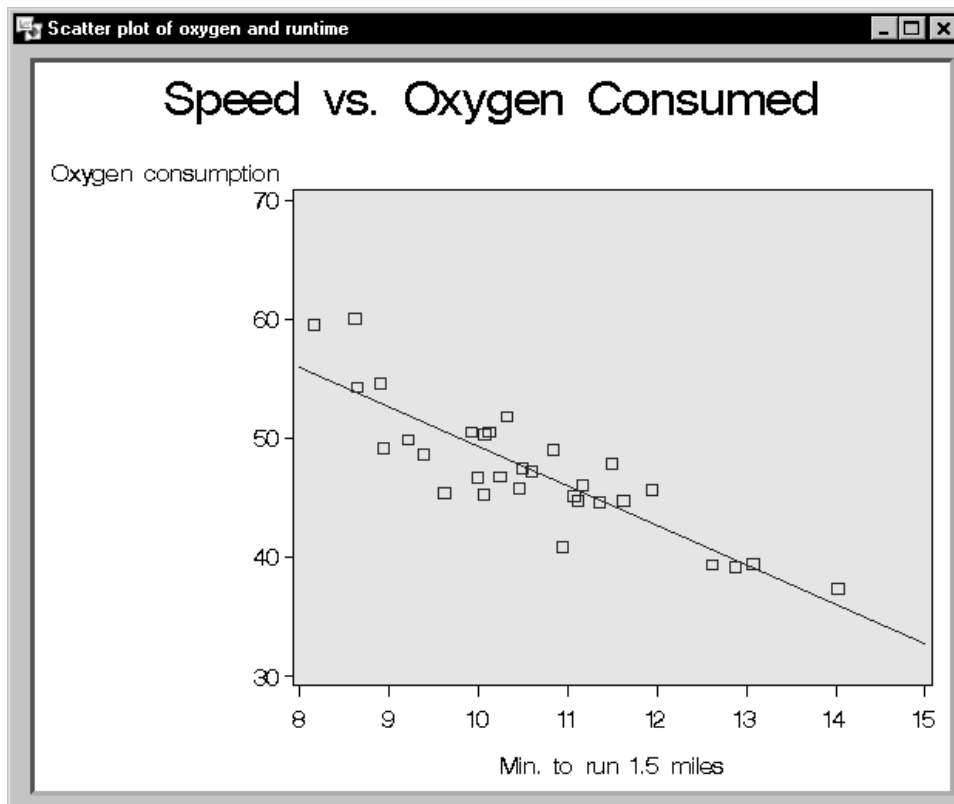


Close the Analysis window to see the project tree. In addition to output, a scatter plot and the SAS code used to perform the regression and create the scatter plot are displayed as nodes in the project tree by default.



**Figure 1.7.** Results in Project Tree

Double-click on the **Scatter plot** node to view the scatter plot that you have created.

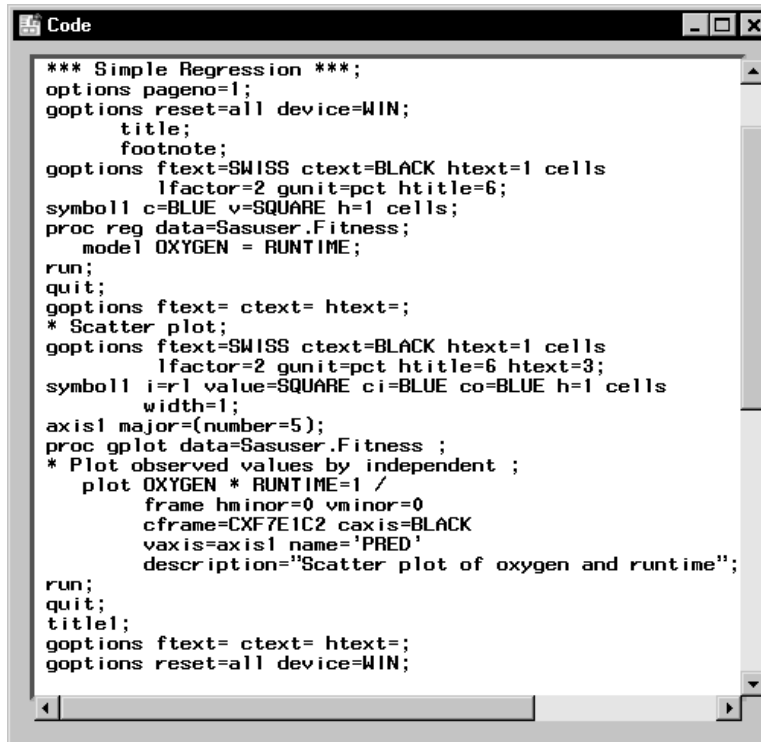


**Figure 1.8.** Scatter Plot

The scatter plot illustrates the results of your simple linear regression: higher oxygen consumption rates are associated with lower running times. You can change the graph to fit the window, edit the graph, or save it to a different format, such as GIF, by selecting **File** → **Save As . . .**

Close the Scatter Plot window to view the **Code** node in the project tree.

Double-click on the **Code** node to view the SAS code that was used to perform the simple linear regression and create the scatter plot.



```

*** Simple Regression ***;
options pageno=1;
goptions reset=all device=WIN;
    title;
    footnote;
goptions ftext=SWISS ctext=BLACK htext=1 cells
    lfactor=2 gunit=pct htitle=6;
symbol1 c=BLUE v=SQUARE h=1 cells;
proc reg data=Sasuser.Fitness;
    model OXYGEN = RUNTIME;
run;
quit;
goptions ftext= ctext= htext=;
* Scatter plot;
goptions ftext=SWISS ctext=BLACK htext=1 cells
    lfactor=2 gunit=pct htitle=6 htext=3;
symbol1 i=r1 value=SQUARE ci=BLUE co=BLUE h=1 cells
    width=1;
axis1 major=(number=5);
proc gplot data=Sasuser.Fitness ;
* Plot observed values by independent ;
    plot OXYGEN * RUNTIME=1 /
        frame hminor=0 vminor=0
        cframe=CXF7E1C2 caxis=BLACK
        vaxis=axis1 name='PRED'
        description="Scatter plot of oxygen and runtime";
run;
quit;
title1;
goptions ftext= ctext= htext=;
goptions reset=all device=WIN;

```

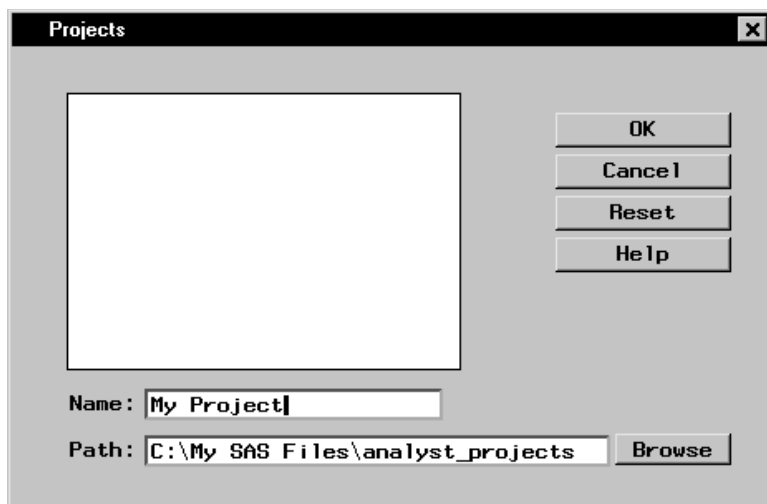
Figure 1.9. Code

---

## Saving a Project

To save the project, follow these steps:

1. Select **New Project** at the top of the project tree.
2. Select **Save. . .** from the pop-up menu.
3. Type **My Project** in the **Name:** field.



**Figure 1.10.** Saving a Project

4. Click **OK**.

---

## Projects

A project is a collection of results from analyses performed on one or more data sets.

A project is displayed as a project tree that contains folders of the different data tables, reports, code, and other results that are associated with the project. Results are presented as nodes in this tree. The folder for each data set contains the results for that data set.

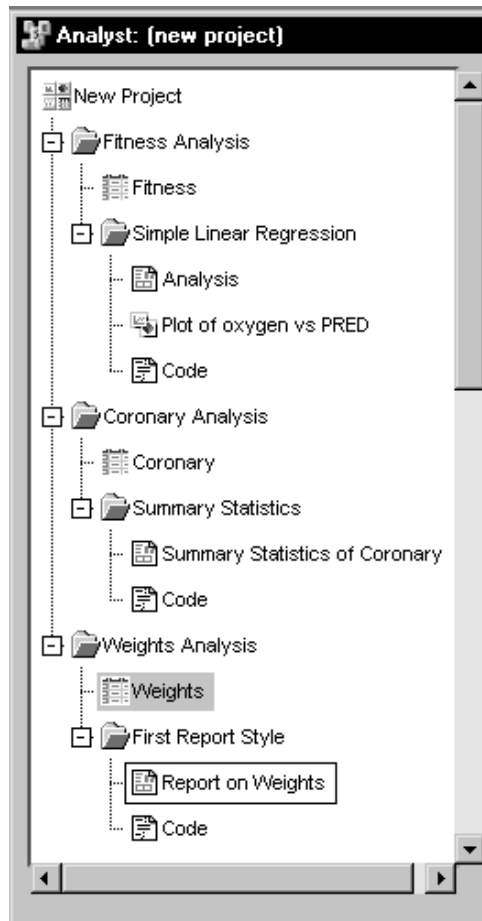


Figure 1.11. Project Tree

---

## Project Folders

You can open and close a folder by clicking the plus (+) or minus (-) sign next to it, by double-clicking on the folder, or by selecting the folder and selecting **Expand** or **Collapse** from the pop-up menu.

---

## Nodes

You can view a node within a folder by double-clicking on the node, or by selecting the node and selecting **View** from the pop-up menu. You can view the same node in a new window by selecting the node and selecting **View in new Window** from the pop-up menu.

From the pop-up menu, you can also delete, print, save, and, except in the case of data tables, rename the node.

If the node is a table, you can view the table in a new window by selecting **View** from the pop-up menu, or you can open the table for analysis by selecting **Open** from the pop-up menu. Also, you can select **Interactive Analysis** from the pop-up menu to invoke SAS/INSIGHT software (if it is installed) to perform interactive exploratory analyses.

---

## Data Table

When you open a data file or SAS library member in Analyst, the data are brought into a data table where you can view and edit the data, perform numerous data transformations, and create new variables.

You can save your data by overwriting your original data source, or you can create a new data table by combining, summarizing, transposing, or taking samples of existing data tables.

The screenshot shows the Analyst software interface. On the left is a project tree with the following structure:

- New Project
  - Fitness Analysis
    - Fitness
    - Simple Linear Regression
      - Analysis
      - Plot of oxygen vs PRED
      - Code
  - Coronary Analysis
    - Coronary
    - Summary Statistics
      - Summary Statistics of Coronary
      - Code
  - Weights Analysis
    - Weights
    - First Report Style
      - Report on Weights
      - Code

On the right is a data table titled "Weights (Browse)". The table has the following data:

	subj	program	strength	time
1	1	CONT	85	1
2	1	CONT	85	2
3	1	CONT	86	3
4	1	CONT	85	4
5	1	CONT	87	5
6	1	CONT	86	6
7	1	CONT	87	7
8	2	CONT	80	1
9	2	CONT	79	2
10	2	CONT	79	3
11	2	CONT	78	4
12	2	CONT	78	5
13	2	CONT	79	6
14	2	CONT	78	7
15	3	CONT	78	1
16	3	CONT	77	2
17	3	CONT	77	3
18	3	CONT	77	4
19	3	CONT	76	5
20	3	CONT	76	6
21	3	CONT	77	7
22	4	CONT	84	1
23	4	CONT	84	2
24	4	CONT	85	3
25	4	CONT	84	4

Figure 1.12. Data Table

## Using the Mouse

You can use the mouse to open project nodes and to select variables in Analyst.

---

## Opening Nodes

Double-click on a node in the project tree to display its contents. Double-clicking on a data set node displays a view of the data set. To open the data set into the data table, select the data set node and select **Open** from the pop-up menu.

---

## Selecting and Removing Variables

In a task dialog, you can select one or more variables for analysis.

To select variables for analysis, double-click on each variable name or highlight the names and click on the appropriate analysis button. To select more than one contiguous variable, press the Shift key while clicking the mouse on the first and last variable that you want to select. All the variables between the first and last variables will be automatically selected. To select noncontiguous variables, press the Ctrl key while clicking the mouse on each variable.

To remove variables from a variable list, double-click on each variable name, or select the variables and click on the **Remove** button.

A C in front of a variable name indicates that it is a character variable.

---

## Accessing Tasks and Help

You can access tasks and help in Analyst by using the menus, the index, and the toolbar on Windows, or the toolbox on other operating systems.

---

## Menus

You can use the menus in Analyst to accomplish a variety of tasks. Click on the pull-down menus at the top of the window, and select items with the mouse, or click the right mouse button within a window to display a pop-up menu. Most items are available from both the pop-up and pull-down menus.

- From the **File** menu, you can access and save projects and data sets, and print reports.

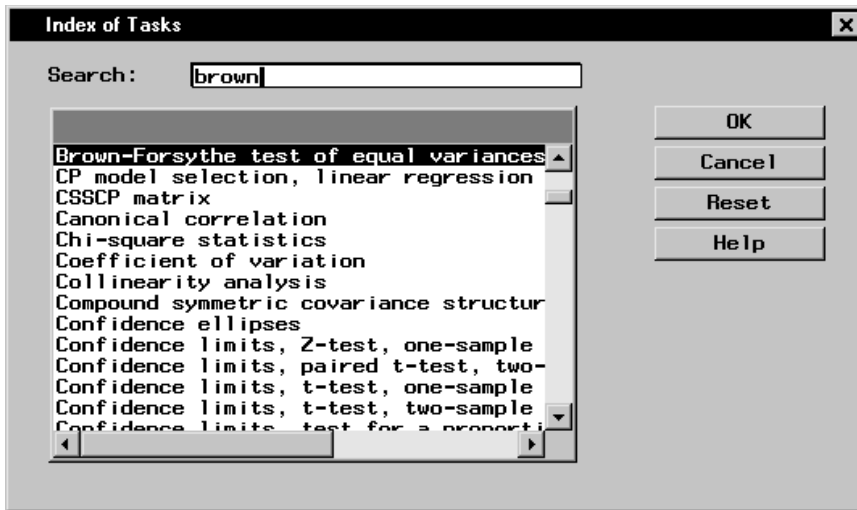


- From the **Edit** menu, you can switch data between Browse and Edit mode, and add, duplicate, and delete columns and rows.
- From the **View** menu, you can move and hide columns, and view the attributes of the data table.
- From the **Tools** menu, you can set the titles for your results, create sample data sets, specify viewer and graph preferences, and assign a new SAS library.
- From the **Data** menu, you can filter, sort, summarize, concatenate, merge, transpose, and apply calculations to your data.
- From the **Reports** menu, you can create detailed and summary reports.
- From the **Graphs** menu, you can generate charts, plots, and histograms.
- From the **Statistics** menu, you can choose statistical analyses and use the index to search for tasks or statistics.
- From the **Help** menu, you can display help for Analyst and the rest of the SAS System.

---

## Index

Use the index to access statistical and graphical tasks through commonly used terms. Select **Statistics** → **Index** . . . to display the Index of Tasks dialog. Click on a term to open a task, or type a term in the **Search:** field to find it in the list.



**Figure 1.13.** Index

For example, if you want to use the Brown-Forsythe test, click on **Brown-Forsythe test of equal variances**, click **OK**, and the One-Way ANOVA task is opened. The Brown-Forsythe test is available in the Tests dialog.

## Toolbar

You can select an Analyst task from the toolbar or toolbox. Click on the icon for a task to select it. By default, the toolbar displays a range of tasks, from opening a file to performing a linear regression. You can display a description for each icon by dragging the mouse cursor over the toolbar. You can also add tasks to the toolbar. See [Chapter 17, “Details,”](#) for more information.



**Figure 1.14.** Toolbar

---

## Getting Help

You can get help in Analyst in three ways:

- the **Help** menu
- the **Help** button on a dialog
- the **Help** icon in the toolbar

### Help Menu

When Analyst is open, select **Help** → **Using This Window** to display help on Analyst. From the main window, **Using This Window** displays the table of contents for Analyst help. From other windows, **Using This Window** displays the help for that particular window. If you are on the Windows operating system, you can go to another help topic through the table of contents or the index.

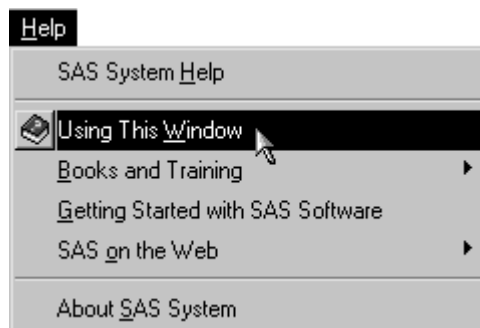



Figure 1.15. Help Menu

### Help Button in Dialogs

Each dialog in Analyst has a **Help** button that you can click on to display the help for that task.

### Help Icon on the Toolbar

Click on the **Help** icon  on the toolbar to display the help table of contents for Analyst.

---

## Graphs

Analyst enables you to create several different kinds of graphs:

- bar charts
- pie charts
- histograms
- box plots
- probability plots
- scatter plots
- contour plots
- surface plots

Use the **Graphs** menu to select the type of graph you want to create.

You can apply settings to all graphs that you produce with Analyst by selecting **Graph Settings . . .** from the **Tools** menu.

---

## Reports

In Analyst, you can create a simple listing of your data or a summary report.

---

### Listing Reports

Select **Reports** → **List Data . . .** to produce a listing of your data.

---

## Summary Reports

You can create a summary report in any one of five table styles. Select **Reports** → **Tables** . . . and select one of the styles that are illustrated.

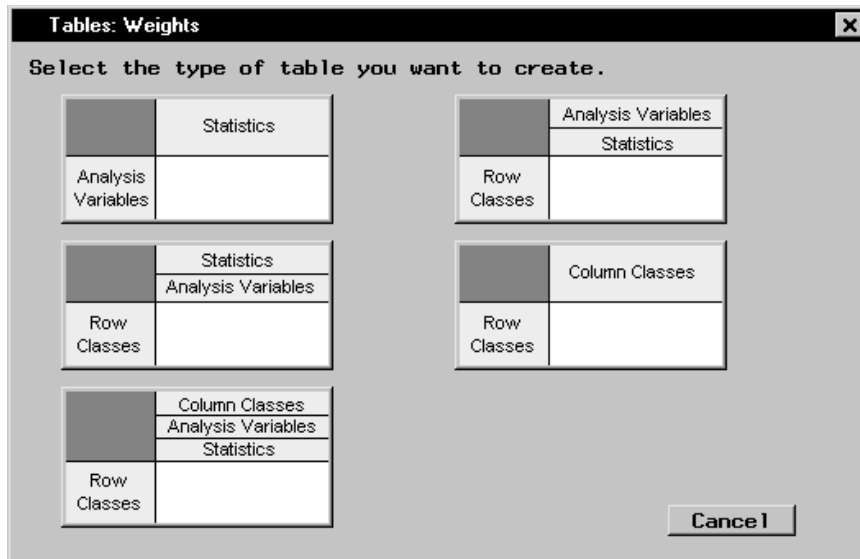


Figure 1.16. Table Styles

---

## Statistical Tasks

Analyst contains a wide range of statistical tasks. You can compute descriptive statistics, perform simple hypothesis tests, and fit models with analysis of variance and regression analysis. There are also tasks for survival analysis, mixed models, repeated measures analysis, and multivariate techniques. Analyst also provides basic sample size and power computations. Graphics are included in most analytical tasks, and you can request many types of graphs directly from the **Graphs** menu.

---

## Descriptive

The descriptive statistical tasks that you can perform on your data include

- summary statistics
- distributions
- correlations
- frequency counts

---

## Table Analysis

In the Table Analysis task, you can create and analyze 2-way to  $n$ -way frequency tables.

---

## Hypothesis Tests

The hypothesis tests that you can perform on your data include

- one-sample Z-test for a mean
- one-sample  $t$ -test for a mean
- one-sample test for a proportion
- one-sample test for a variance
- two-sample  $t$ -test for means
- two-sample paired  $t$ -test for means
- two-sample test for proportions
- two-sample test for variances

---

## ANOVA

You can perform one-way, nonparametric one-way, and factorial analysis of variance (ANOVA). You can also fit the general linear model, perform repeated measurements ANOVA, and fit basic mixed models.

---

## Regression

The Linear Regression task provides linear and multiple linear regression analysis. The Simple Linear Regression task predicts a dependent variable from a single independent quantitative variable.

The Logistic Regression task investigates the relationship between a binary outcome (such as success and failure) or an ordinal outcome (such as mild, moderate, and severe) and a set of explanatory variables.

---

## Multivariate

The Principal Components task computes principal components from a set of variables.

The Canonical Correlation task describes the relationship between two sets of variables by finding a small number of linear combinations from each set of variables that have the highest possible between-set correlations.

---

## Survival

The Life Tables task computes nonparametric estimates of the survival distribution of data that may be right censored due to withdrawals or study termination. This task computes rank tests and a likelihood ratio test for testing homogeneity of survival functions across strata.

The Proportional Hazards task performs regression analysis of survival data based on the Cox proportional hazards model.

---

## Sample Size

The Sample Size tasks enable you to determine the power of a test, given the sample size, or the sample size required to obtain a specified power. These calculations can be made for a variety of situations, including  $t$ -tests, confidence intervals, tests of equivalence, and one-way ANOVA. These are prospective power and sample size computations; retrospective power computations are provided for some of the analytical tasks.





# Chapter 2

## The Data Table

### Chapter Contents

---

<b>Introduction</b> . . . . .	27
<b>Bringing in Data</b> . . . . .	28
Opening Local Files . . . . .	28
Opening SAS Files . . . . .	29
Using the Query Window . . . . .	29
<b>Modifying Tables</b> . . . . .	32
Viewing and Editing Data . . . . .	33
Working with Columns . . . . .	33
Working with Rows . . . . .	40
Typing in Data Values . . . . .	42
The Data Menu . . . . .	42
Computing New Variables . . . . .	43
Recoding Ranges . . . . .	44
Computing Log Transformations . . . . .	47
Generating Random Variates . . . . .	47
Combining Tables . . . . .	48
Splitting Columns . . . . .	51
Subsetting Data . . . . .	52
Example: Modifying a Data Table . . . . .	53
<b>Saving and Exporting Data</b> . . . . .	64
Saving Data . . . . .	64
Saving Data to a SAS Library . . . . .	64
Reserved Names . . . . .	65

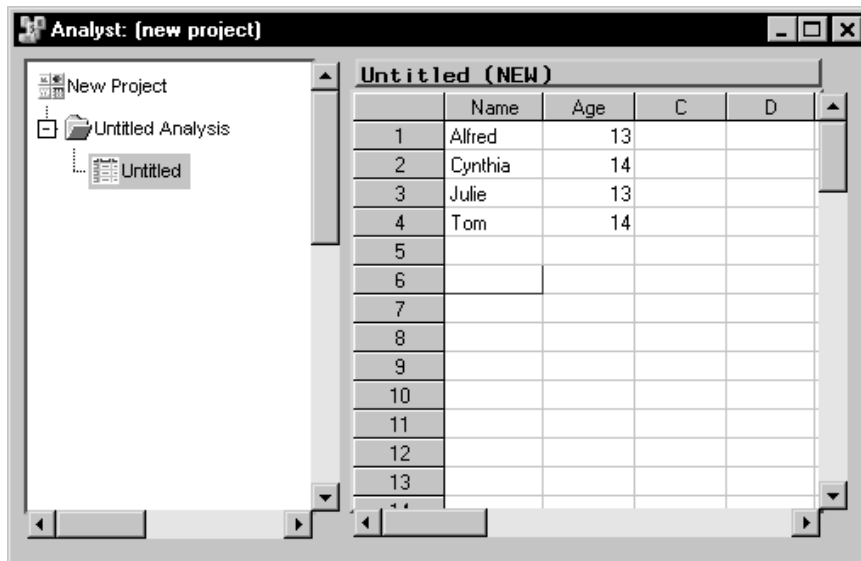
Exporting Data to Different File Formats . . . . . 66

# Chapter 2

## The Data Table

### Introduction

The Analyst data table provides a spreadsheet view of your data set, where rows correspond to observations and columns correspond to variables. You can type data directly into the table as well as display data from SAS data sets, data views, and other sources. You can also customize the appearance of the data table by rearranging rows and columns, changing column formats, and applying filters.



**Figure 2.1.** The Data Table

You can enter data into the data table by typing values directly into table cells. In a new table, the first value you enter in a column determines the column type. That is, if the first value you type is numeric, then the column

is defined as numeric and no longer permits character values. Once you have entered data into the data table, you can immediately generate graphics and perform analyses. However, you must save the new table as a data set before you can subset, sort, and transform your data.

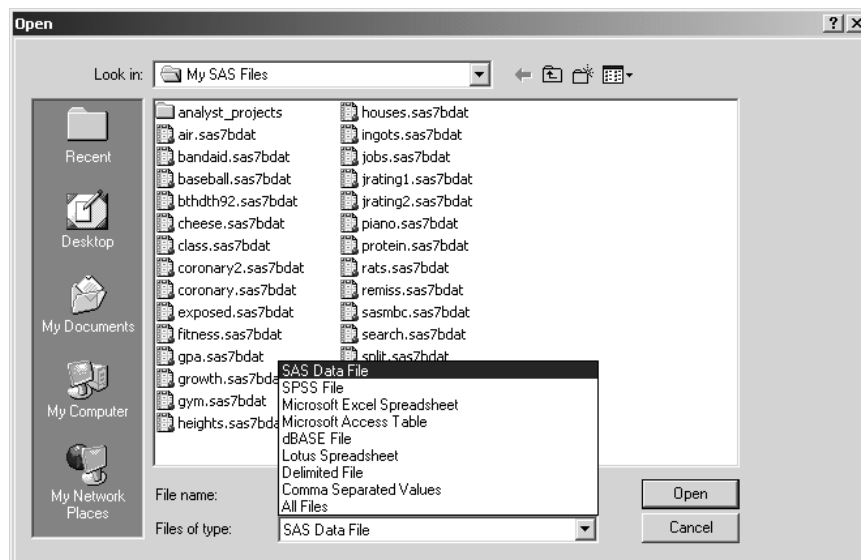
---

## Bringing in Data

---

### Opening Local Files

The Analyst Application supports many different file formats, including SAS data sets, Excel spreadsheets, Lotus spreadsheets, SPSS portable files, and delimited files. You can open data files from your operating system's directories or folders and bring them into the data table by selecting **File** → **Open** . . .



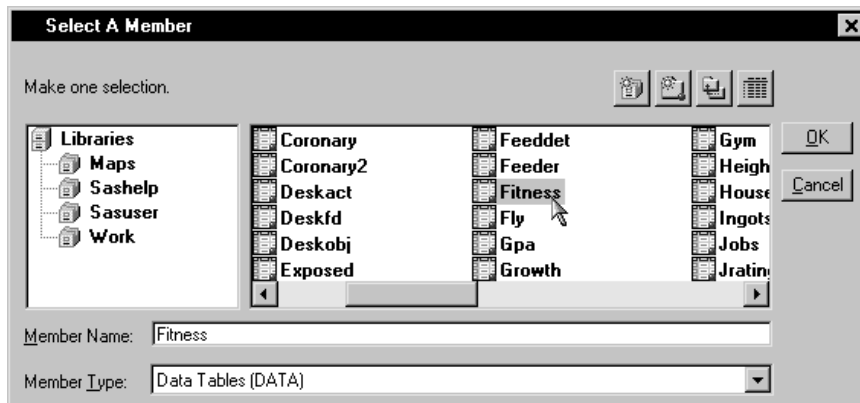
**Figure 2.2.** Open Dialog

In the Open dialog, select a file and click **Open** to bring the contents of the file into the data table. External files (files that were not created in SAS)

opened into Analyst are converted into SAS data sets. The source files are not altered.

## Opening SAS Files

You can bring SAS data sets or data views into the Analyst data table by selecting **File** → **Open By SAS Name** . . .



**Figure 2.3.** Select A Member Dialog

Select a SAS library from the list of **Libraries** and select a member. Click **OK** to bring the contents of the SAS data set or data view into the data table.

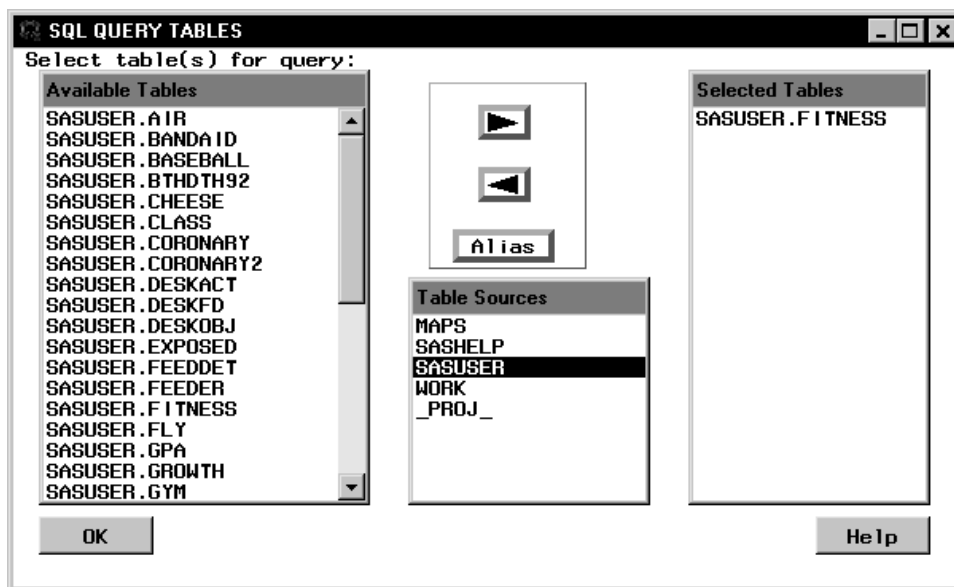
## Using the Query Window

You can use the Query window to reduce the number of variables that you load into the data table. You can also use the Query window to bring more than one data set into the data table, as well as write SQL queries to filter the data.

### Opening a New Query

You can use the Query window to bring selected columns of data from one or more SAS data sets into the data table. The Query window opens a view of the data set that cannot be edited. You can, however, save the view as a

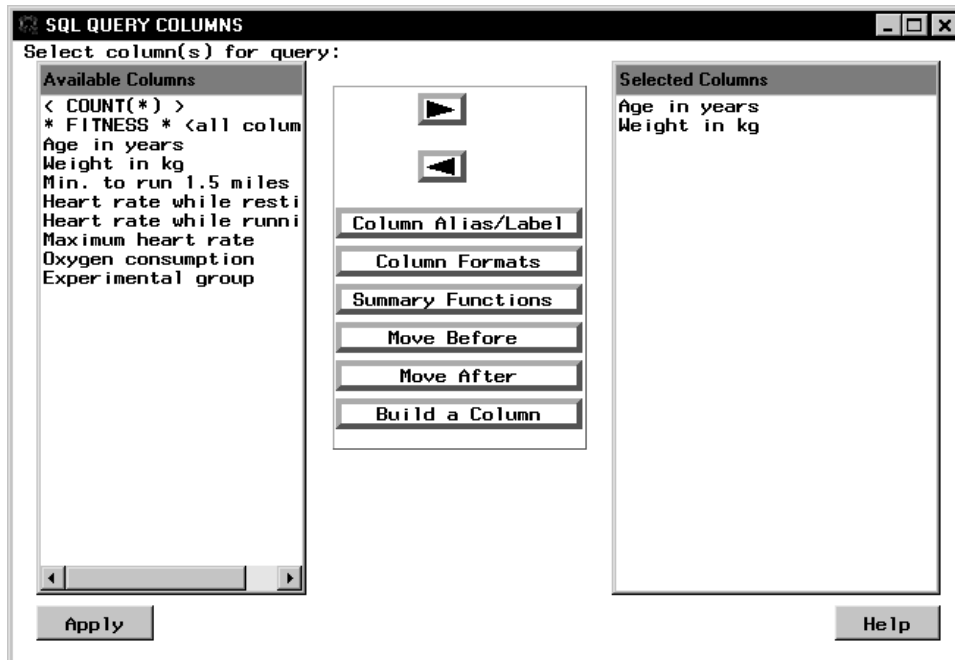
SAS data set that you can edit. To save the view as a SAS data set, select **File** → **Save As By SAS Name . . .**



**Figure 2.4.** SQL QUERY TABLES Window

Select **File** → **Open With New Query . . .** to open the SQL QUERY TABLES window. Select one or more tables to use in your query and click on the right arrow.

Click **OK** to display the SQL QUERY COLUMNS window. Select the columns that you want to include in the query and click on the right arrow.



**Figure 2.5.** SQL QUERY COLUMNS Window

Select **File** → **Close** to exit the Query window and open the data view into the Analyst data table.

The query is added as a node to your project tree, and the selected columns are brought into the data table. The name of the query node is generated by Analyst in the form QUERY $nnnn$ .

**Caution:** If you select the Analyst window while in the Query window, the resulting query is not returned to Analyst.

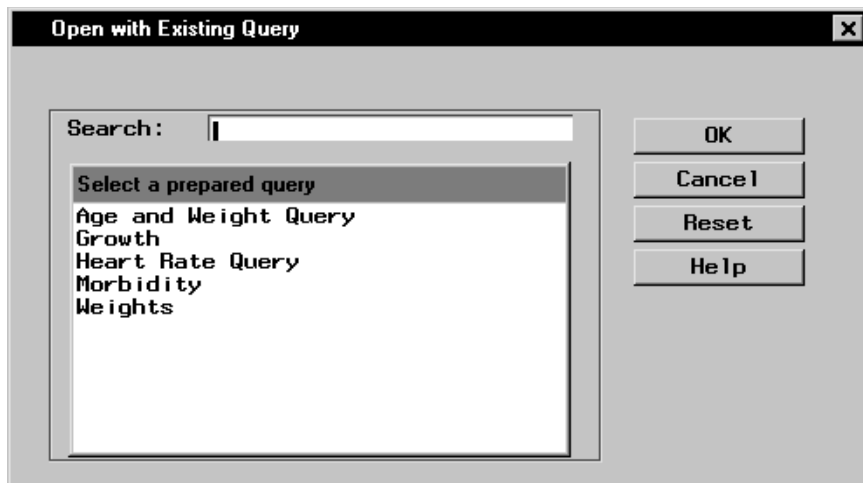
### ***Saving and Opening an Existing Query***

Once you have used the Query window to create views of SAS data, you can bring these views into Analyst.

To create a query to use later, prepare your query in the Query window, and select **File** → **Save Query** → **Save as QUERY to Include later** in the SQL

QUERY COLUMNS window. Select the SAS library, catalog, and library member name.

To open a saved query in Analyst, select **File** → **Open With Existing Query** ... The Open with Existing Query window searches for saved queries in all available SAS libraries.



**Figure 2.6.** Open with Existing Query Window

You can also use the Query window to apply an SQL query to your data. Refer to the Query window documentation for more information.

---

## Modifying Tables

When you have brought your data into the Analyst data table, you can change the organization and apply calculations to the data. You must be in Edit or Shared Edit mode to make modifications to the data table.



---

## Viewing and Editing Data

To prevent changes to a table while you are viewing it, select **Edit** → **Mode** → **Browse**.

To make changes to the table, select **Edit** → **Mode** → **Edit**. While you are in Edit mode, no one else is able to make changes to the table.

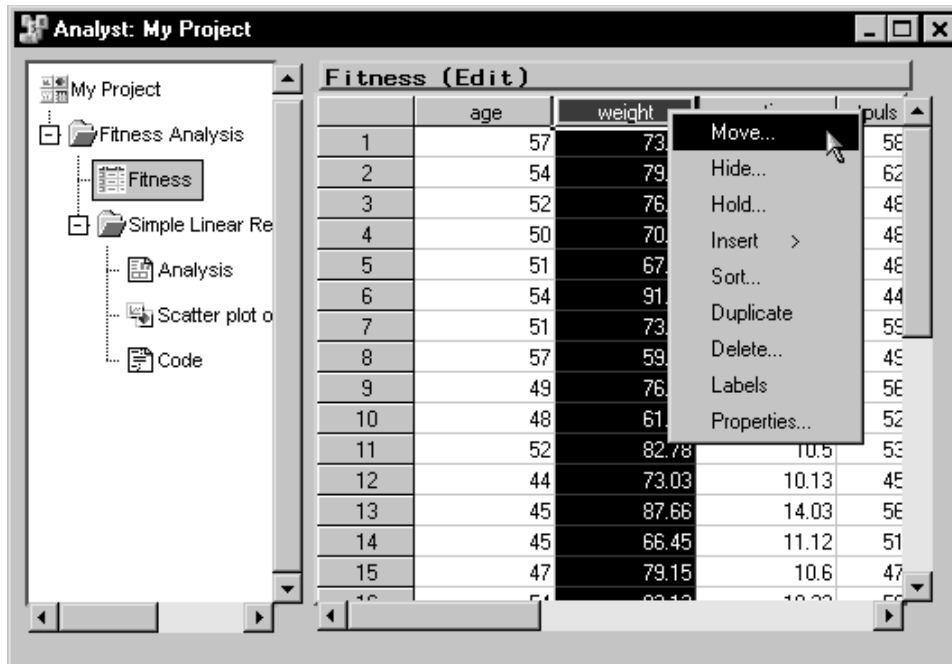
To allow more than one person to make concurrent changes to the table, select **Edit** → **Mode** → **Shared Edit**. The record you are editing is locked while you are editing it, but other users can make changes to other records in the table.

When you are in Edit or Shared Edit mode, you can make changes to the data table by selecting a cell and typing in it.

---

## Working with Columns

You can perform several operations on data table columns by selecting items from a pop-up menu. To display the pop-up menu for a column, select the column and click the right mouse button.

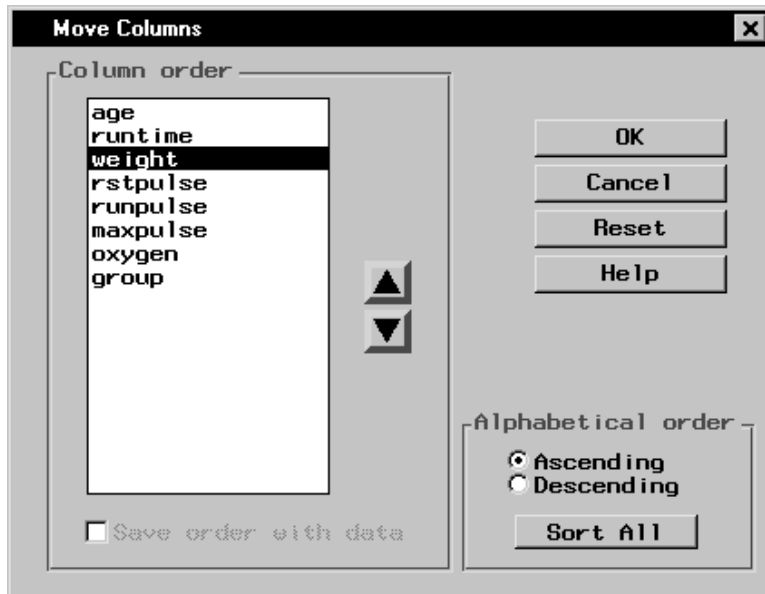


**Figure 2.7.** Column Pop-up Menu

These items are also available from the **View**, **Edit**, and **Data** menus.

### ***Moving Columns***

You can move columns by selecting one or more columns and selecting **Move . . .** from the pop-up menu to display the Move Columns dialog.



**Figure 2.8.** Move Columns Dialog

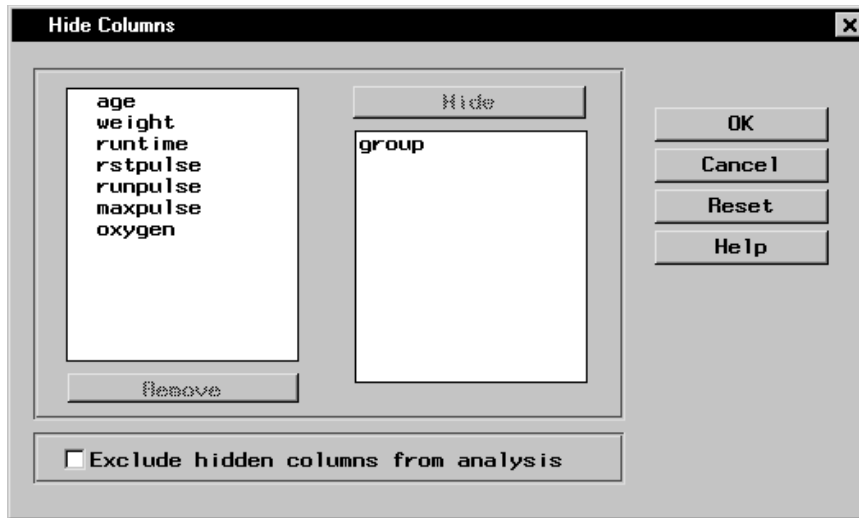
To move a column, select it in the **Column order** list, then click on the arrows to move it to the appropriate spot. Sort the columns by selecting **Ascending** and **Descending** under the **Alphabetical order** heading. Click on the **Sort All** button to sort the columns.

Select **Save order with data** to save this order with the data file. You must be in Edit mode to save the order with the data file.

Click **OK** when the columns are in the desired order.

### **Hiding Columns**

To hide a column or columns from displaying in the data table, select the columns and select **Hide ...** from the pop-up menu to display the Hide Columns dialog. Hidden columns are still used in an analysis unless you specify that they be excluded.



**Figure 2.9.** Hide Columns Dialog

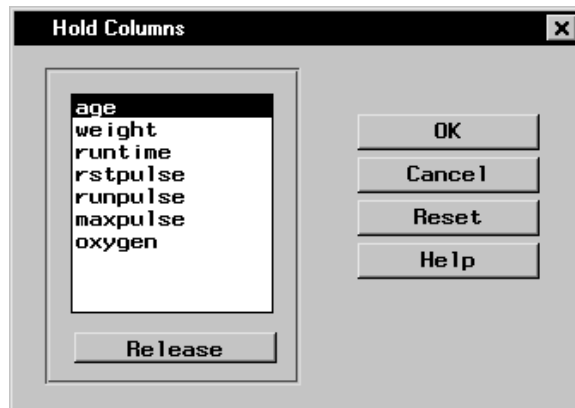
To hide columns, select the desired columns and click on the **Hide** button.

To unhide columns, select the desired columns and click on the **Remove** button.

Select **Exclude hidden columns from analysis** to specify that the hidden columns be unavailable for Analyst tasks.

### **Holding Columns**

To hold a column and all the columns to the left of it in place while you scroll through the columns in the data table, select a column, and select **Hold . . .** from the pop-up menu to display the Hold Columns dialog.



**Figure 2.10.** Hold Columns Dialog

Select a column from the column list and click **OK** to hold it.

Select a held column from the column list and click on the **Release** button to release it.

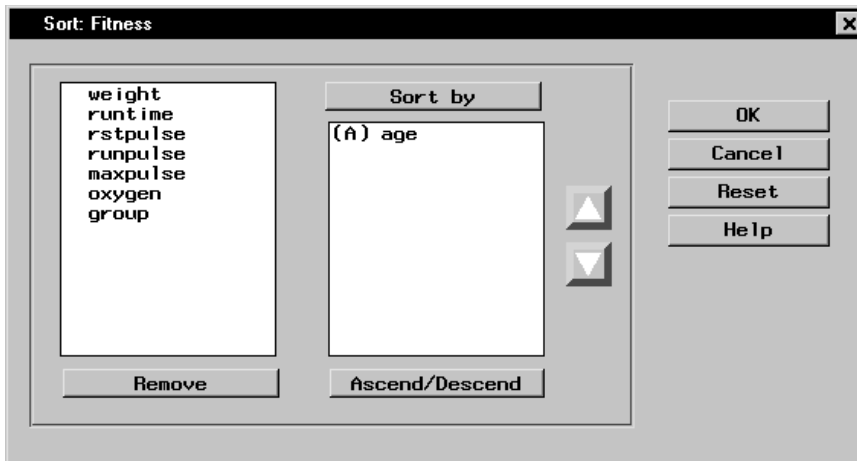
### **Inserting Columns**

To insert one or more columns, select a column and select **Insert** from the pop-up menu. Then select the column type **Character** or **Numeric**. The new column is inserted to the left of the selected column. If you select more than one column, columns equal to the number you have selected are inserted to the left of the first column. If no column is selected, the new column is added to the end of the table.

You must be in Edit mode to insert columns.

### **Sorting Columns**

Select a column and select **Sort . . .** from the pop-up menu to display the Sort dialog. Sort the rows in the data table by the selected column's values.



**Figure 2.11.** Sort Dialog

Select columns from the candidate list and click on the **Sort by** button to specify the column values to use in sorting.

Use the up and down arrows next to the **Sort by** list to specify the desired column sort order.

Select a variable in the **Sort by** list and click on the **Ascend/Descend** button to sort the rows in the data table in ascending or descending alphabetical order of column values. The rows are sorted in ascending order by default. You must be in Edit mode to sort columns.

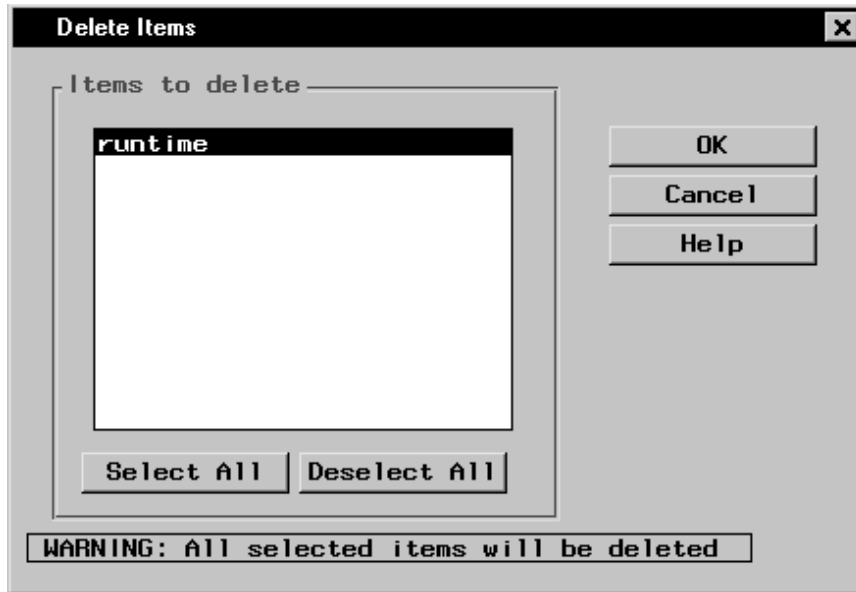
### **Duplicating Columns**

To duplicate one or more columns, select a column and select **Duplicate** from the pop-up menu. The duplicated column is inserted to the left of the selected column. If you select more than one column, each column is duplicated to the left of the first selected column.

You must be in Edit mode to duplicate columns.

## Deleting Columns

To delete a column, select the column and select **Delete . . .** from the pop-up menu to display the Delete Items dialog.



**Figure 2.12.** Delete Items Dialog

Select the columns that you want to delete and click **OK**. To avoid deleting any columns, deselect all columns or click on the **Cancel** button.

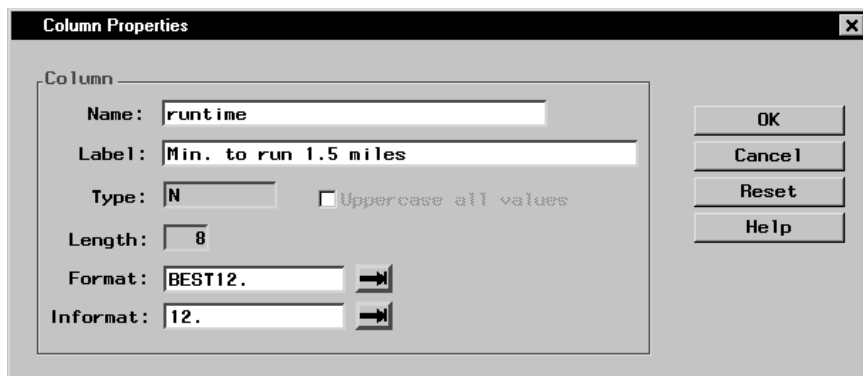
You must be in Edit mode to delete columns.

## Displaying Variable Labels

You can switch between displaying variable names as column headings in the data table and displaying labels as column headings in the data table by selecting a column and selecting **Labels** from the pop-up menu.

## Column Properties

Select a column and select **Properties . . .** from the pop-up menu to display the Column Properties dialog.



**Figure 2.13.** Column Properties Dialog

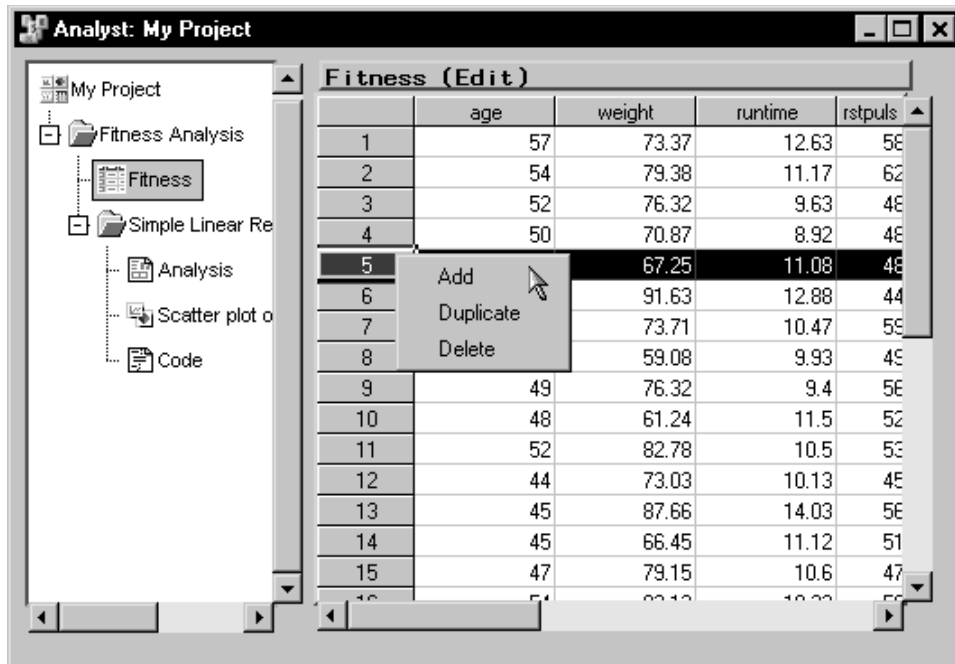
The Column Properties dialog displays the name, label, type (numeric or character), length, format, and informat of the selected column. If the data table is in edit mode, you can change the name, label, format and informat for the variable that the column represents. Otherwise, you can only view the information.

---

## Working with Rows

You can add, duplicate, and delete rows. To display the pop-up menu for a row, select the row and click the right mouse button.





**Figure 2.14.** Row Pop-up Menu

These items are also available from the **Edit** menu.

### **Adding a Row**

To add a row to the end of the table, select a row and select **Add** from the pop-up menu.

You must be in Edit or Shared Edit mode to add a row.

### **Duplicating a Row**

To duplicate a row, select the row, and select **Duplicate** from the pop-up menu.

You must be in Edit or Shared Edit mode to duplicate a row.

### Deleting a Row

To delete a row, select the row, and select **Delete** from the pop-up menu.

You must be in Edit or Shared Edit mode to delete a row.

---

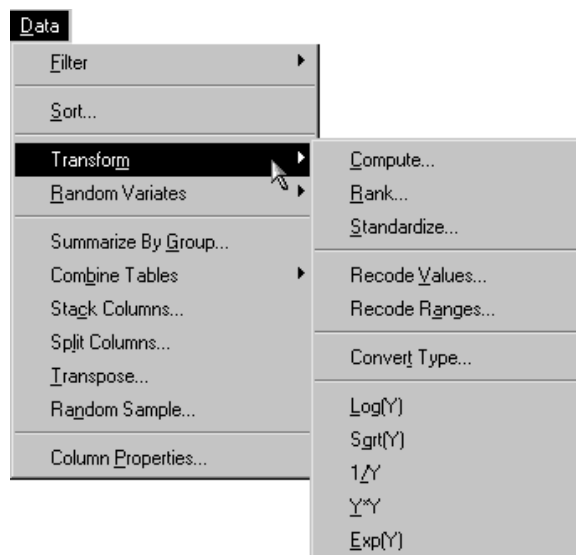
## Typing in Data Values

You can change the data in a cell by selecting the cell and typing in the new value.

---

## The Data Menu

From the **Data** menu, you can filter, sort, summarize, concatenate, merge, transpose, and apply calculations to your data.



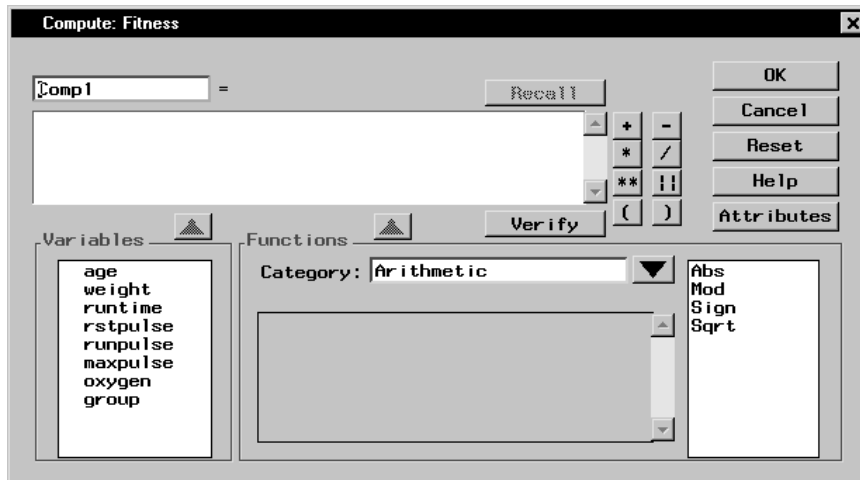
**Figure 2.15.** Data Menu

The following topics describe a few important **Data** menu tasks. Two other important **Data** menu tasks, stacking columns and recoding values, are described and used in Chapter 16. **Data** menu tasks not described in this book include ranking and standardizing data, converting the values of a variable from numeric to character or character to numeric, producing a summary data set, transposing a data set, taking a random sample, and creating a new

column that is a square, square root, reciprocal, or exponential of an existing column. Consult the Analyst online help for more information about these tasks.

## Computing New Variables

You can specify an expression for creating a new column in the data table. Select **Data** → **Transform** → **Compute . . .** to display the Compute dialog.



**Figure 2.16.** Compute Dialog

Type the expression in the box under the new column name, or use a combination of typing and selecting variables, functions, and operators. A numeric column is created by default.

Click on an operator at the right of the expression box to add it to the expression. You can also type in an operator.

To add a variable to the expression, double-click on the variable name or select it and click on the arrow above the **Variables** list. You can also type in a variable name.

Functions are organized into categories. Select a category by clicking on the arrow next to the **Category:** field. Review information about a function by

selecting it. This information appears in the box to the left of the function list. Add a function to the expression by double-clicking on it or selecting the function and clicking on the arrow above the **Functions** box. You can also type in any SAS function. The functions displayed are a subset of all SAS functions.

By default, the column name is *CompN*, where *N* is the lowest number that produces a unique name. Replace the default column name by typing in one of your choosing.

The **Attributes** button displays the Column Attributes dialog, in which you can specify the name, label, and other attributes for your computed column. If you want to create a column with character values, use this dialog to set the variable type to character. Numeric is the default variable type.

Click on the **Verify** button to make sure your expression is valid. Function parameters are not verified, and the variable type is not taken into account.

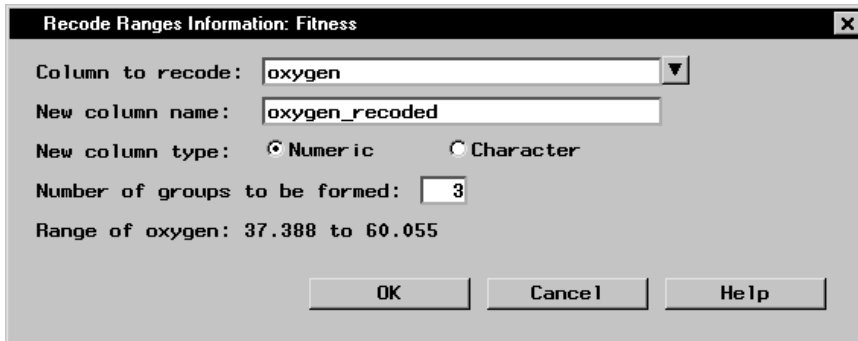
If you have already used the Compute dialog to add a column to the current data table, click on the **Recall** button to fill the expression box and the Column Attributes dialog with the most recent expression and attributes.

---

## Recoding Ranges

In performing an analysis, you may want to work with a particular factor as a classification variable rather than as a continuous variable. Recoding ranges enables you to create a new variable with discrete levels based on the ranges of values of an existing variable.

Select **Data** → **Transform** → **Recode Ranges . . .** to designate the column whose ranges you want to use.



**Figure 2.17.** Recode Ranges Information Dialog

Click on the arrow next to **Column to recode:** to select a numeric column from the current data table.

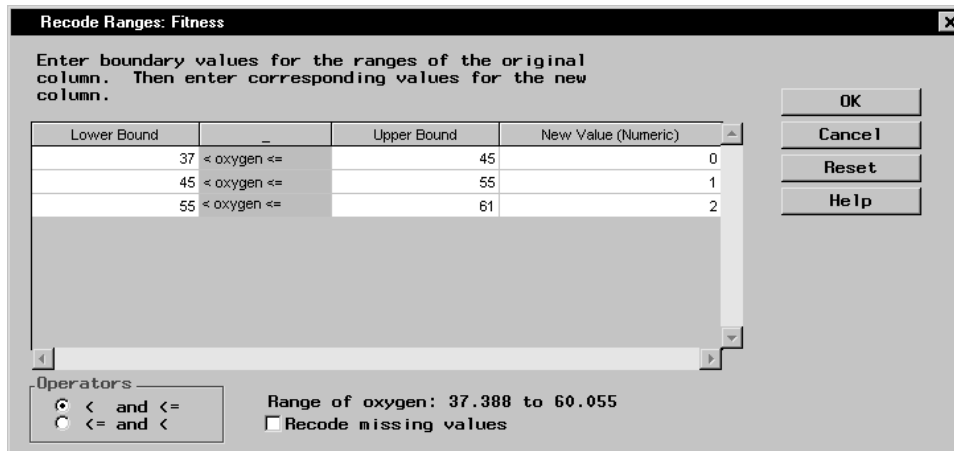
Specify the name of the new column that will contain the new data values. The new column has a default name, which you can type over with a name of your choosing.

The new column type can be character or numeric. If you select **Character**, you can use a character string to correspond to each range.

You must specify the number of groups that the current range will be divided into.

To help you decide how many groups to form, the range of the existing column is displayed at the bottom of this dialog.

After you have selected a column to recode and the number of groups that you want the new variable to have, click **OK** to display a dialog in which you can specify the recoding to be performed.



**Figure 2.18.** Recode Ranges Dialog

Use this dialog to substitute new values for the original ranges of the column specified in the Recode Ranges Information dialog. The number of rows in the table corresponds to the number of groups.

The **Lower Bound** is the lower boundary of a range. The **Upper Bound** is the upper boundary of a range. The upper boundary is automatically transferred to the next range's lower boundary. Only the first  $N - 1$  cells of the **Upper Bound** need to be filled in.

Type in a character or numeric value to correspond to the range. If you do not type in a value, a missing value (blank) is assigned to the range.

Under **Operators**, you can control what happens to column values that fall on a range boundary. The first option groups these values with smaller values; the second option groups these values with larger values.

If you select **Recode missing values** and the lowest lower bound is left blank, missing values are placed in the lowest new group. If you don't select **Recode missing values**, missing values remain missing.

The range of the existing column is displayed at the bottom of this dialog.

---

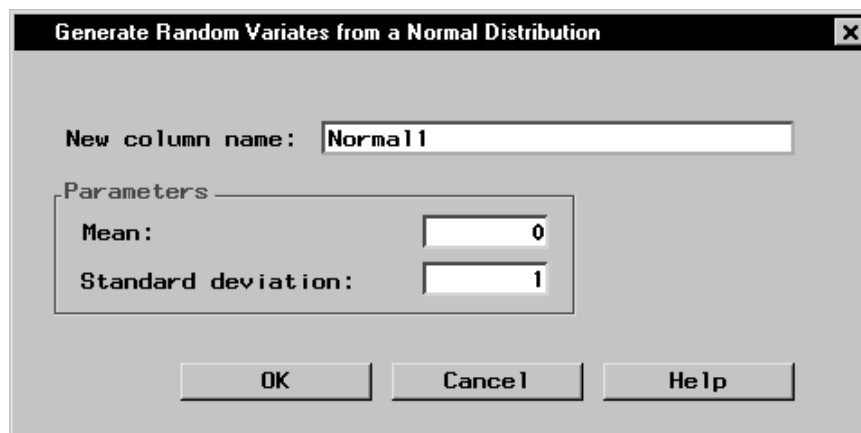
## Computing Log Transformations

Select a column and select **Data** → **Transform** → **Log(Y)** to calculate the natural logarithm of the values in the selected column. A new column containing the logarithm of each value is created. Other transformations, such as exponentiating and taking a square root, are also available from the **Transform** item in the **Data** menu.

---

## Generating Random Variates

To generate random variates, select **Data** → **Random Variates**, and then select the distribution to be used for generating the random variates.



**Figure 2.19.** Generate Random Variates from a Normal Distribution Dialog

You can leave the new column name as the default or specify a new column name in the **New column name:** field.

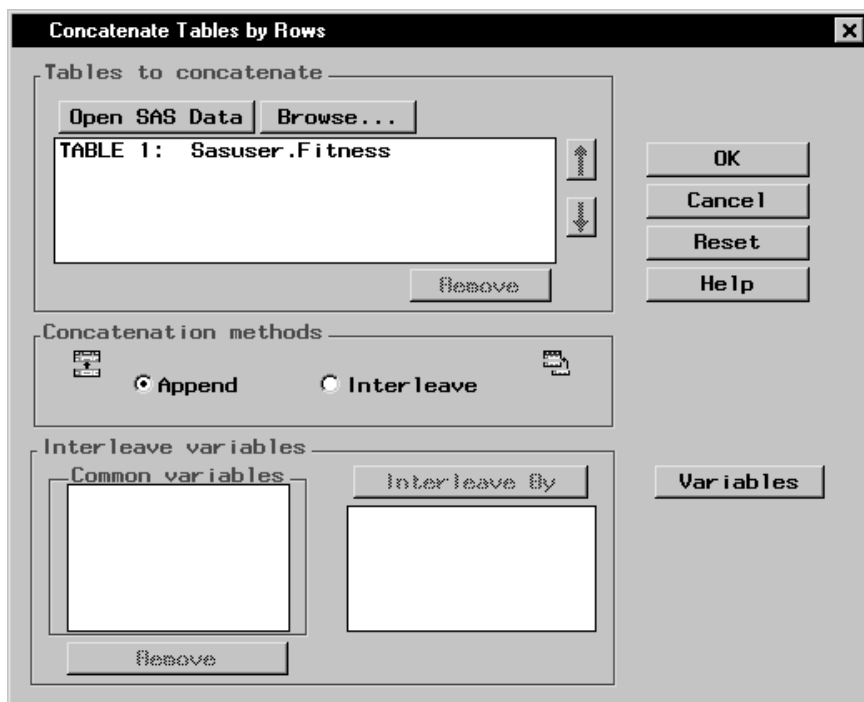
Enter a value for each parameter. Click **OK** to create a column with the specified distribution.

## Combining Tables

You can concatenate the rows or merge the columns from two or more tables.

### Concatenating Tables by Rows

To vertically join tables by concatenating their rows, select **Data** → **Combine Tables** → **Concatenate By Rows** . . .



**Figure 2.20.** Concatenate Tables by Rows Dialog

Click on the **Open SAS Data** button to open SAS data tables. Click on the **Browse** button to select a file from your operating system's directory.

To change the order of the tables that you are appending, select a table and click on the up or down arrow to move the table one level up or one level down in the list.



To remove a table from the list, select the table and click on the **Remove** button.

Select **Append** to append the tables that you have selected. If you have chosen to append the tables, you can change the order of tables in the list. When you append tables, the rows of the first table are followed by the rows of the succeeding tables.

Select **Interleave** to interleave the rows of the tables.

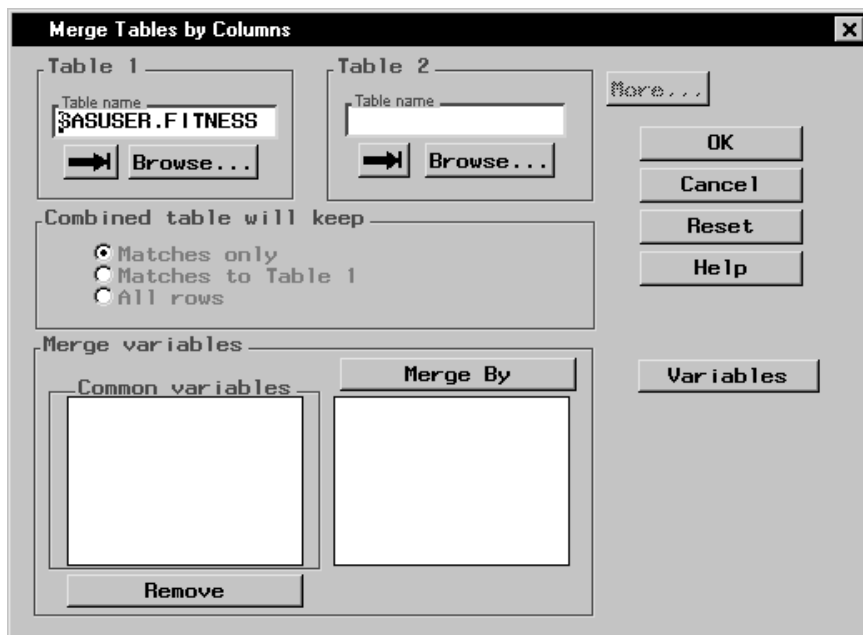
Common variables among the tables you have chosen to concatenate are listed in the **Common variables** list. Select a common variable and click on the **Interleave By** button to add it to the list of variables to interleave by. When you interleave table rows, the rows of the table are combined and ordered according to the common variables that you have selected.

Select a variable and click on the **Remove** button to remove it from the list of **Interleave By** variables.

Click on the **Variables** button to choose the variables that you want to keep in your concatenation. By default, when you concatenate by rows, the resulting table contains only the common variables.

### ***Merging Tables by Columns***

To join tables horizontally by merging their columns, select **Data** → **Combine Tables** → **Merge By Columns . . .**



**Figure 2.21.** Merge Tables by Columns Dialog

In the Merge Tables by Columns dialog, you can select data tables to merge and the variables you will keep in the merged table. You can merge up to six tables. Type the name of the table in the **Table name** field, click on the arrow to select a SAS data table, or click on the **Browse** button to select a file from a directory.

Click on the **More** button to merge more than two tables.

You can choose whether the new combined table displays only matching rows, rows that match those in **Table 1**, or all rows.

Common variables among the tables you have chosen to combine are listed in the **Common variables** list.

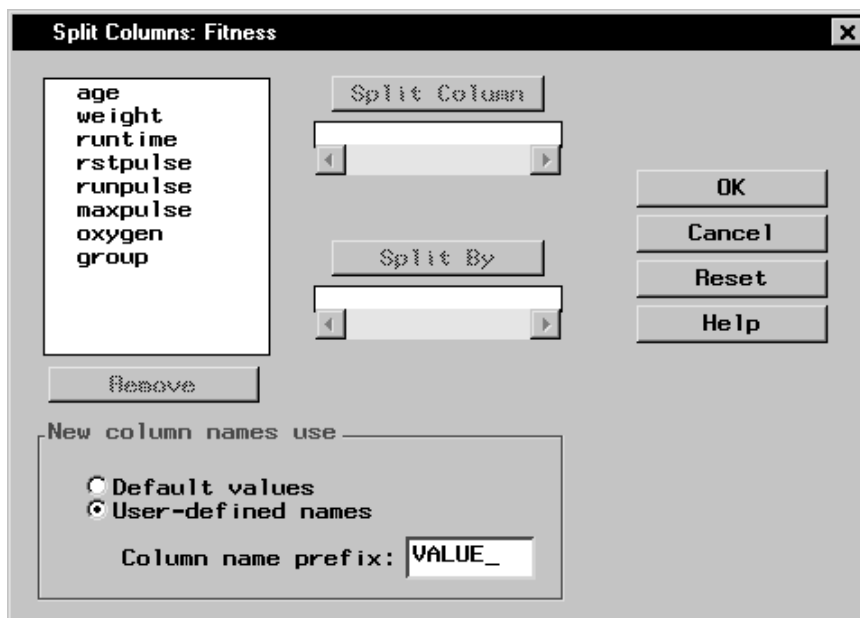
Select a common variable and click on the **Merge By** button to add it to the list of variables to combine the tables by.

Select a variable and click on the **Remove** button to remove it from the list of **Merge By** variables.

Click on the **Variables** button to choose the variables that you want to keep in your merged table. By default, when you merge by columns, the resulting table contains all the variables.

## Splitting Columns

You can split selected columns to output a new column whenever the value of a variable changes. Select **Data** → **Split Columns** . . . to display the Split Columns dialog.



**Figure 2.22.** Split Columns Dialog

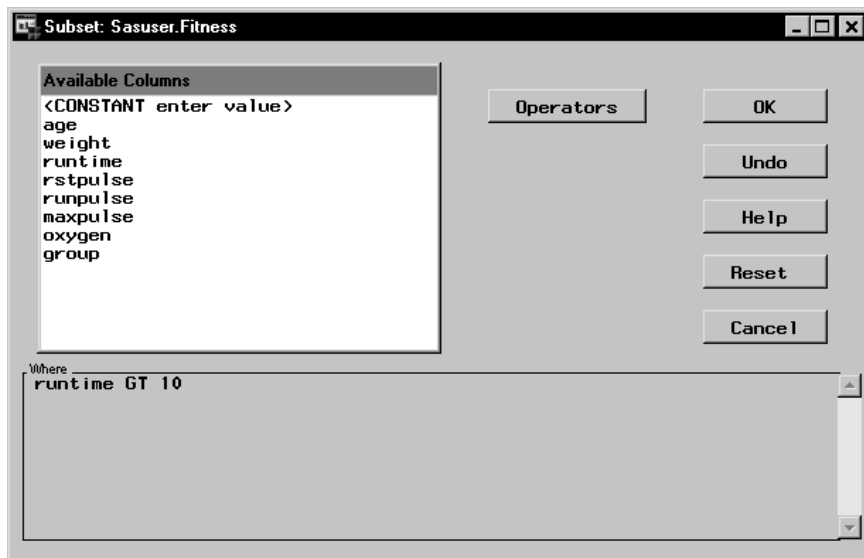
Select a column from the candidate list and click on the **Split Column** button to designate a column to split.

Select a variable from the candidate list and click on the **Split By** button to designate a variable to split the first column by.

You can use the default names or type in new names for the split column if the type of the **Split By** column is character. Numeric columns do not have default names.

## Subsetting Data

You can view a subset of your data by selecting **Data** → **Filter** → **Subset Data . . .** In the Subset dialog, you can apply a Where clause to your data.



**Figure 2.23.** Subset Dialog

All subsequent analyses are run on the subset of the data.

Select **Data** → **Filter** → **None** if you do not want to subset your data, or if you want to remove an existing subset. **None** is the default.

To save the subsetted data, select **File** → **Save As . . .** If you select **File** → **Save**, the entire data set, and not just the subset, is saved.

---

## Example: Modifying a Data Table

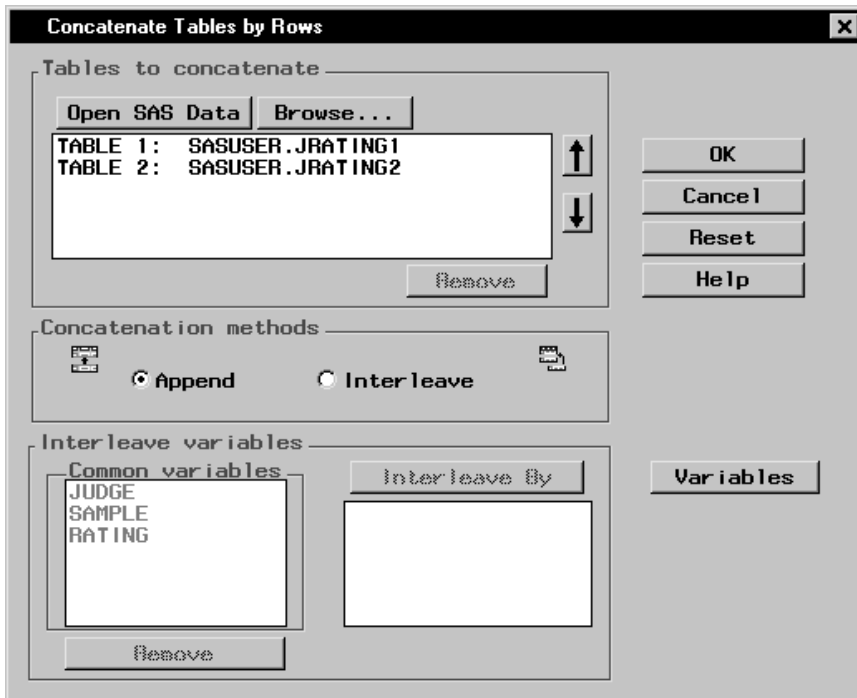
In this example, you combine selected columns from two data sets and edit them in a new data table. This example assumes that you have no data set loaded in the Analyst data table. If you do, select **File** → **New** before starting the example.

Each data set contains the results of taste tests of breakfast cereal. Each cereal is rated by several judges, on a scale of 1 to 5. After you concatenate the two data sets, you split the rating column by sample number.

### *Open Data Sets for Editing*

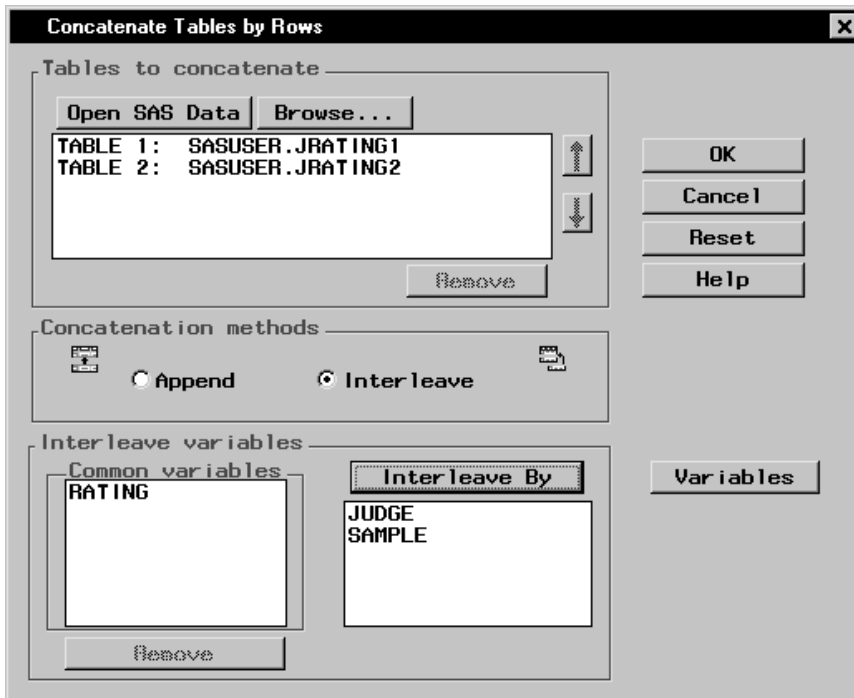
To select the data sets and bring them into a new Analyst data table, follow these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select JRating1 and JRating2.
3. Click **OK** to create the sample data sets in your Sasuser directory.
4. Select **Data** → **Combine Tables** → **Concatenate By Rows . . .**
5. Click on the **Open SAS Data** button. Select Sasuser from the list of **Libraries**. Select Jrating1 from the list of members. Click **OK**.
6. In the Concatenate Tables by Rows dialog, click on the **Open SAS Data** button again. Select Sasuser from the list of **Libraries**. Select Jrating2 from the list of members. Click **OK**.



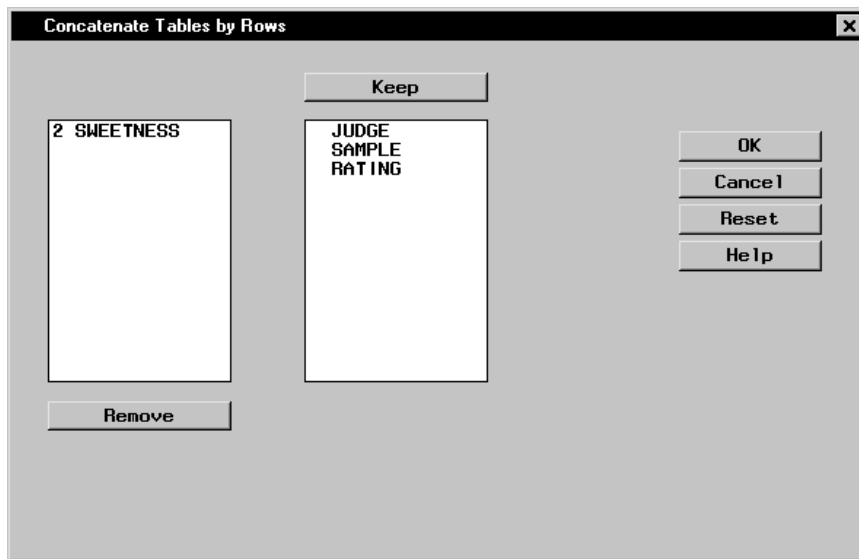
**Figure 2.24.** Concatenate Tables by Rows Dialog

7. Select **Interleave**.
8. Select JUDGE and SAMPLE from the list of **Common variables** and click on the **Interleave By** button to use JUDGE and SAMPLE as the variables by which the rows of the data tables will be combined.



**Figure 2.25.** Interleave by Common Variables

9. Click on the **Variables** button to select the columns to include in the new data table.

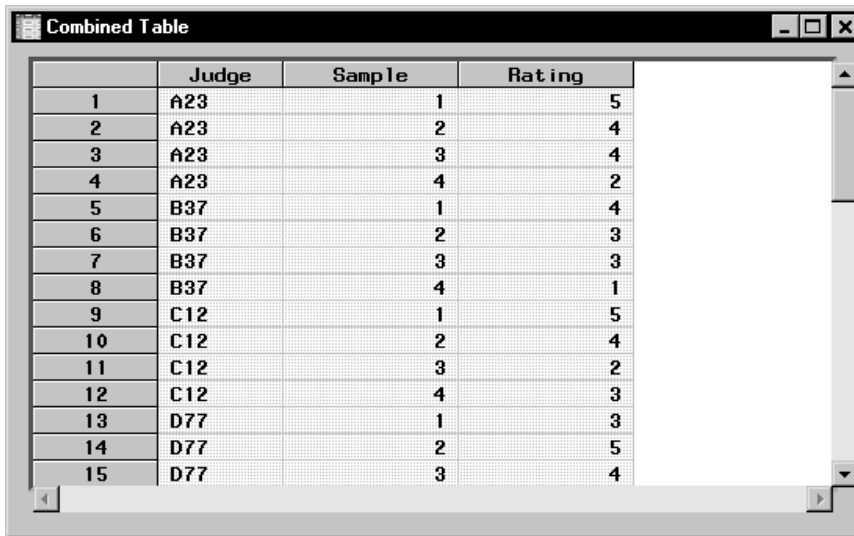


**Figure 2.26.** Selected Columns for New Data Table

Only those columns common to both data tables are kept by default, as shown in the **Keep** list. The column **SWEETNESS** is not kept as part of the resulting table. The number preceding the column name **SWEETNESS** represents the data table to which this variable belongs.

10. Click **OK** to return to the Concatenate Tables by Rows dialog. Click **OK** again to display the new combined data table in a results window.



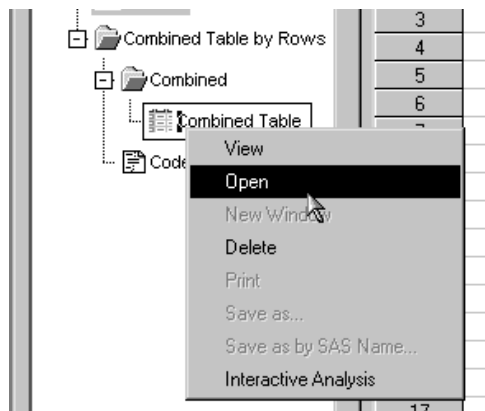


The screenshot shows a window titled "Combined Table" with a table containing 15 rows and 4 columns. The columns are labeled "Judge", "Sample", and "Rating". The first column is an index from 1 to 15. The "Judge" column lists identifiers A23, B37, and C12, with D77 appearing in the last three rows. The "Sample" column lists integers from 1 to 4. The "Rating" column lists integers from 1 to 5. The table is displayed in a standard grid format with a scroll bar on the right side.

	Judge	Sample	Rating
1	A23	1	5
2	A23	2	4
3	A23	3	4
4	A23	4	2
5	B37	1	4
6	B37	2	3
7	B37	3	3
8	B37	4	1
9	C12	1	5
10	C12	2	4
11	C12	3	2
12	C12	4	3
13	D77	1	3
14	D77	2	5
15	D77	3	4

**Figure 2.27.** Combined Table

11. To modify the combined table, you need to open it in the Analyst data table. Close the results window. Select the **Combined Table** node in the project tree and click the right mouse button to display the pop-up menu. Select **Open**.



**Figure 2.28.** Opening the Combined Table

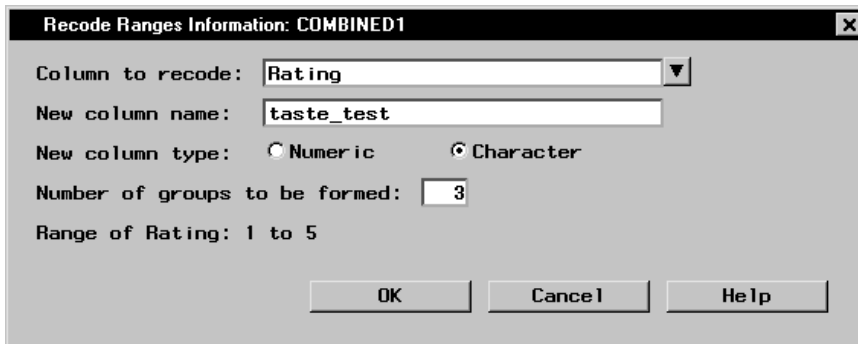
12. By default, data tables are opened in Browse mode. Select **Edit** → **Mode** → **Edit** to change the mode from Browse to Edit.

### **Modify the Data**

In the data table you can modify the data by splitting columns so that a new column is generated when the value of a variable changes. You can also subdivide data into ranges.

To subdivide the data into ranges and split the columns according to sample number, follow these steps:

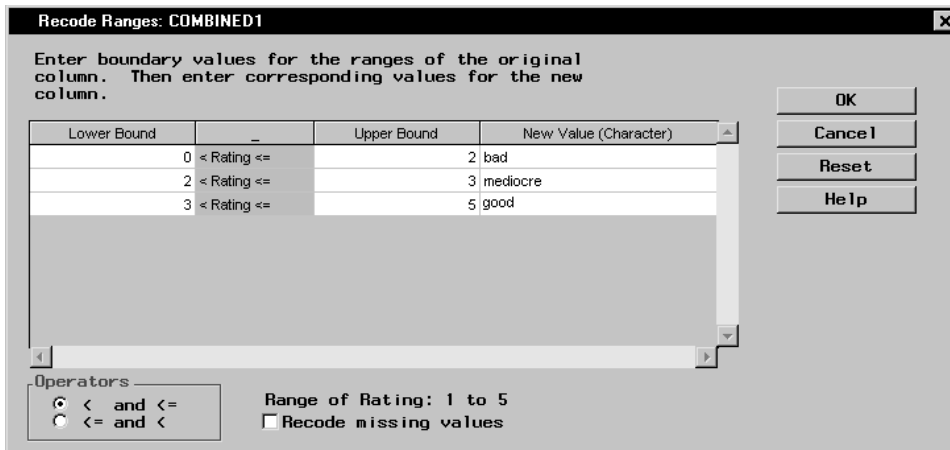
1. Divide the taste test results into three categories: good, mediocre, and bad. Select **Data** → **Transform** → **Recode Ranges . . .**
2. Click on the arrow next to **Column to recode:** and select Rating. Type **taste\_test** in the **New column name:** field. Change **New column type:** to **Character**. Type **3** in the **Number of groups to be formed:** field to designate three taste test ranges.



**Figure 2.29.** Recode Ranges Information Dialog

Click **OK** to specify the new ranges.

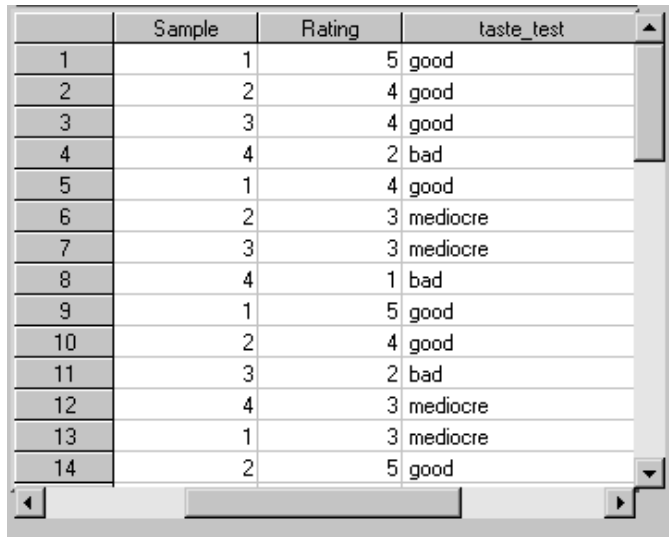
3. In the first row, type **0** in the **Lower Bound** column and **2** in the **Upper Bound** column. Type **bad** in the **New Value** column.
4. When you press the Enter key, the upper bound value of the previous row is automatically filled in as the lower bound of the current row. Type **3** in the **Upper Bound** column and **mediocre** in the **New Value** column.
5. Move your cursor to the third row. Type **5** in the **Upper Bound** column and **good** in the **New Value** column.



**Figure 2.30.** Boundary Values

6. Click **OK** to save your new boundary values.

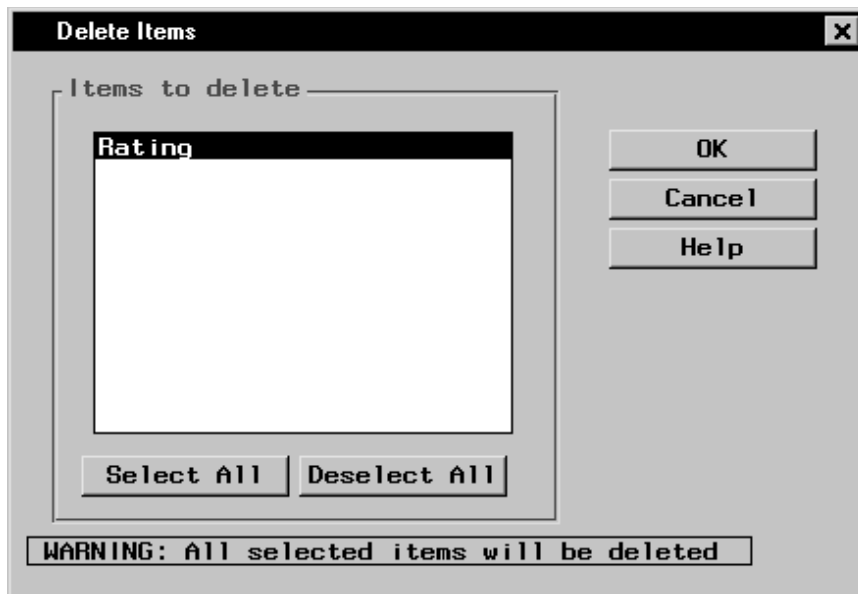
In the new table, the new ranges are displayed in the **taste\_test** column.



	Sample	Rating	taste_test
1	1	5	good
2	2	4	good
3	3	4	good
4	4	2	bad
5	1	4	good
6	2	3	mediocre
7	3	3	mediocre
8	4	1	bad
9	1	5	good
10	2	4	good
11	3	2	bad
12	4	3	mediocre
13	1	3	mediocre
14	2	5	good

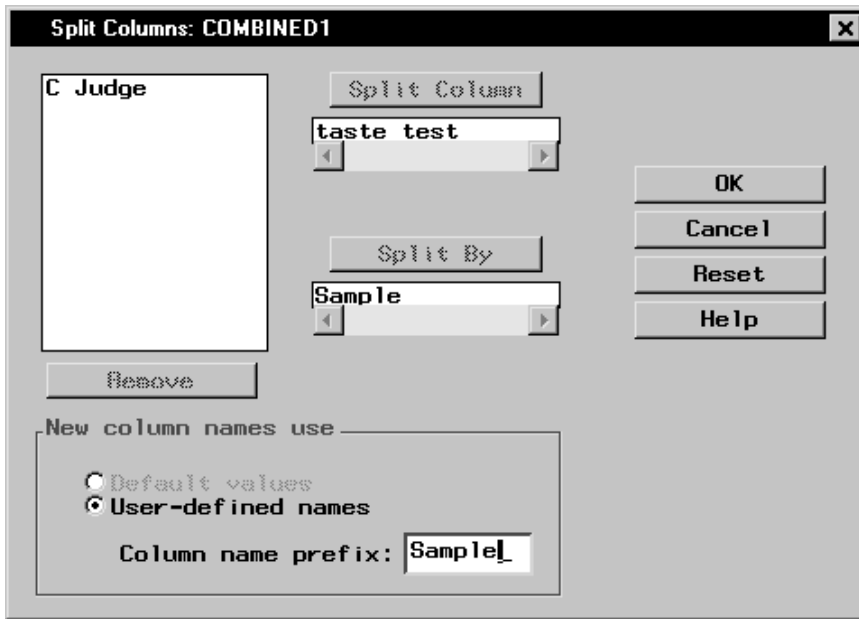
**Figure 2.31.** Table with taste\_test Column

7. Remove the **Rating** column by selecting the column and selecting **Delete . . .** from the pop-up menu. Click **OK** in the Delete Items dialog.



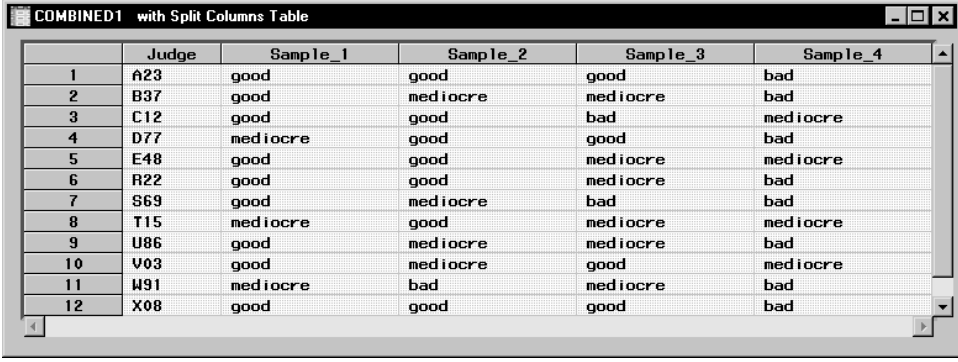
**Figure 2.32.** Delete Rating Column

8. You are going to split the `taste_test` column by the `Sample` column so that a taste test for each sample is displayed by judge. Select **Data** → **Split Columns . . .**
9. In the Split Columns dialog, select `taste_test` from the list and click on the **Split Column** button. Select `Sample` from the list and click on the **Split By** button.
10. Select **User-defined names** for the column names. Type `Sample_` in the **Column name prefix:** field.



**Figure 2.33.** Taste\_test Column Split by Sample

11. Click **OK**. The resulting table displays the results of the taste test by each participating judge.



	Judge	Sample_1	Sample_2	Sample_3	Sample_4
1	A23	good	good	good	bad
2	B37	good	mediocre	mediocre	bad
3	C12	good	good	bad	mediocre
4	D77	mediocre	good	good	bad
5	E48	good	good	mediocre	mediocre
6	R22	good	good	mediocre	bad
7	S69	good	mediocre	bad	bad
8	T15	mediocre	good	mediocre	mediocre
9	U86	good	mediocre	mediocre	bad
10	V03	good	mediocre	good	mediocre
11	W91	mediocre	bad	mediocre	bad
12	X08	good	good	good	bad

Figure 2.34. Split Columns Table

---

## Saving and Exporting Data

---

### Saving Data

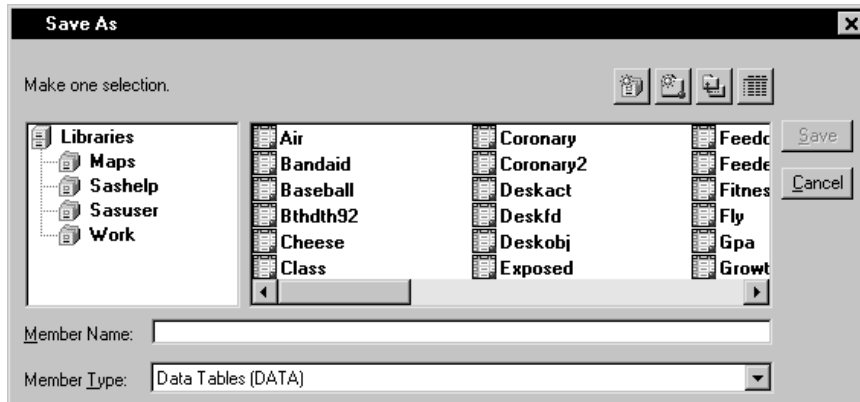
To save changes made to the current data set, select **File** → **Save**.

---

### Saving Data to a SAS Library

Select **File** → **Save As By SAS Name . . .** to save the current table as a SAS data set.





**Figure 2.35.** Save As Dialog

Select a library from the list of **Libraries**. Select an existing data set from the member list or type a member name for the new data set in the field next to **Member Name:**. Click on the **Save** button to save the data set. The new data set is automatically opened into Analyst.

---

## Reserved Names

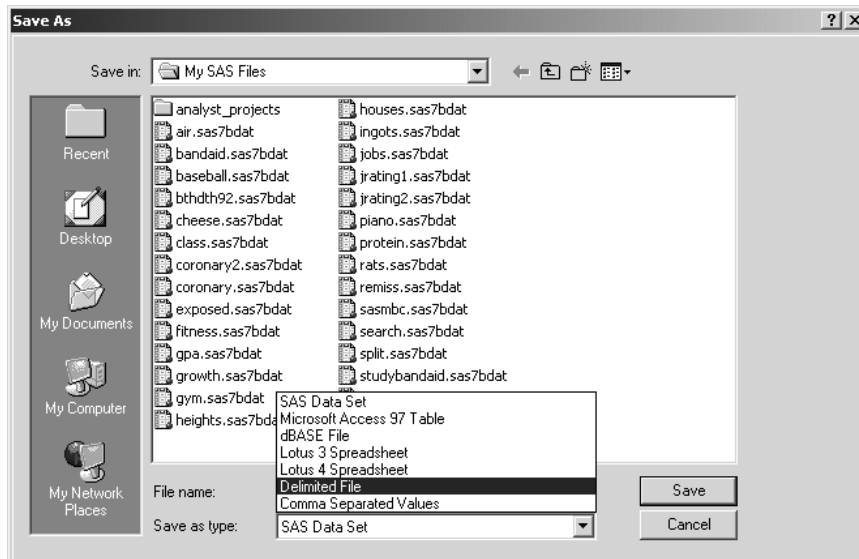
The following names are reserved by Analyst and should not be used to refer to tables.

The `_proj_` libref points to the current project library where project files are stored. This libref is dynamically assigned each time a project is opened.

A `_tmp_` libref is assigned by Analyst as needed. `_tmp_` is also used as the stem of names for temporary data sets used by Analyst, for example, `_tmp_0439`.

## Exporting Data to Different File Formats

You can save files to any export format that is supported by SAS Software on your platform. For example, you can export a SAS data table to a delimited file. Select **File** → **Save As . . .** to export a data table to a different format.



**Figure 2.36.** Save As Dialog

# Chapter 3

## Managing Results in Projects

### Chapter Contents

---

<b>Introduction</b> . . . . .	69
<b>Managing Projects</b> . . . . .	69
Creating a Project . . . . .	69
Saving a Project . . . . .	70
Saving a Project Under Another Name . . . . .	71
Renaming a Folder . . . . .	72
Deleting Nodes from a Project . . . . .	72
Deleting a Project . . . . .	73
Opening Existing Projects . . . . .	73
<b>Using Code</b> . . . . .	73
Viewing Code in the Code Window . . . . .	73
Copying Code to the Program Editor Window . . . . .	74
<b>Printing and Saving Results</b> . . . . .	75
Saving Text Results . . . . .	75
Saving a Graph Result as a File . . . . .	76
Saving a Result as a Catalog Entry . . . . .	77
Printing Results . . . . .	78
<b>Example: Create and Export Histograms</b> . . . . .	78
Open the Project . . . . .	79
Save the Project Under Another Name . . . . .	79
Generate Histograms . . . . .	80
Export Histograms . . . . .	85



# Chapter 3

## Managing Results in Projects

---

### Introduction

An Analyst project is a collection of results from analyses performed on one or more data sets.

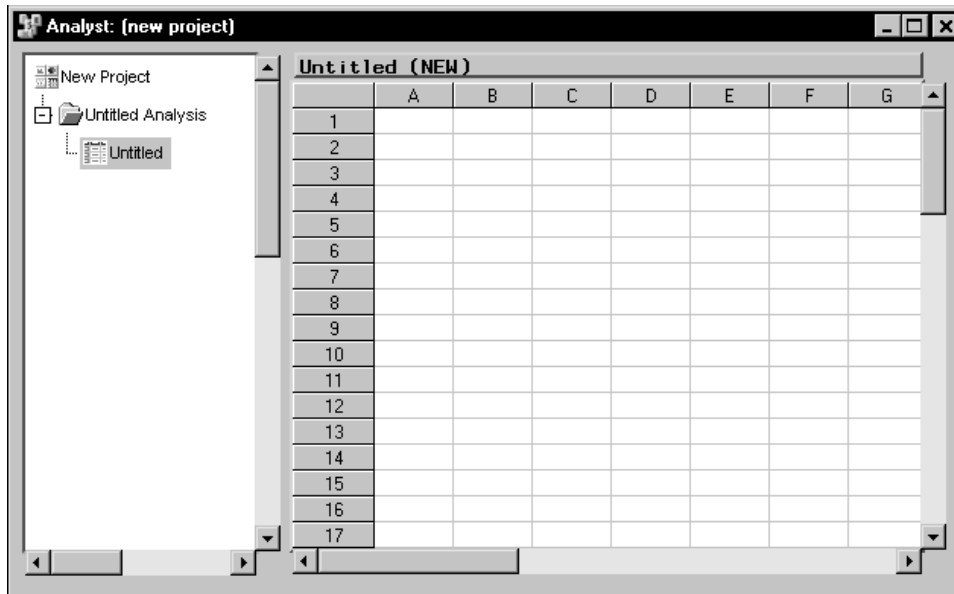
Select **Projects** from the **File** menu to create, open, save, and delete Analyst projects.

---

### Managing Projects

#### Creating a Project

If you do not have any existing projects when you invoke the Analyst application, a new project is automatically created for you. If you already have existing projects, and you want to create a new project, select **File** → **Projects** → **New** to create a new project. A new project tree is displayed.



**Figure 3.1.** New Project

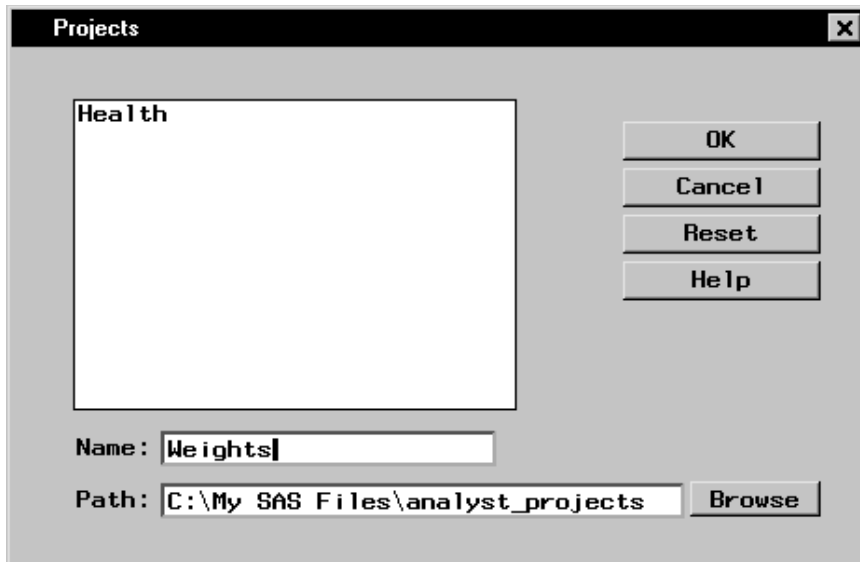
A folder named **Untitled Analysis** that contains a data node named **Untitled** is automatically created in the new project. You can enter data into the data table, open a SAS data file, or open external data files such as Excel files. If you open data into the data table, the folder name is replaced by the name of the data set that you open. If you enter data into the data table, the folder name is replaced when you save the data set.

---

## Saving a Project

To save a project, select **File** → **Projects** → **Save**. A new project must contain a named data table before it can be saved.

When you save a new project, you are prompted to give the project a name.



**Figure 3.2.** Projects Dialog

Type the name of the new project in the **Name:** field. Click on the **Browse** button to search for a directory in which to save the project. Click **OK** to save the project. By default, Analyst projects are saved in the `analyst_projects` directory within the `Sasuser` directory.

---

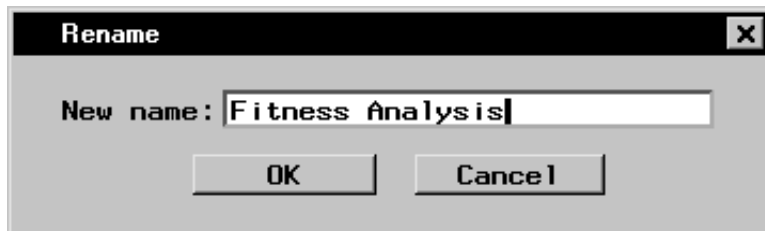
## Saving a Project Under Another Name

To save the contents of a project under another name, select **File** → **Projects** → **Save As...** and type the new name of the project in the **Name:** field. Click on the **Browse** button to search for a directory in which to save the project. Click **OK** to save the project with the new name. The original project, with its original name, still exists.

---

## Renaming a Folder

To rename a folder within a project, select the folder with the right mouse button, and select **Rename . . .** from the pop-up menu.



**Figure 3.3.** Rename Dialog

Type the new name of the folder in the **New name:** field and click **OK**.

---

## Deleting Nodes from a Project

You can delete individual nodes in a project without deleting the project itself. To delete a node, select the node and select **Delete** from the pop-up menu.

Deleting a SAS data set node from the project tree does not delete it from the directory in which it resides. For example, if you open the **Fitness** data set and perform analyses on it, it is not deleted from the **Sasuser** library when you delete it from the project tree.

Deleting an output data set that you have generated from the SAS data set does delete it from the **analyst\_projects** folder where it resides. For example, if you create a data table by combining selected columns from two SAS data sets, the data table that you created is deleted when you remove it from the project tree.



---

## Deleting a Project

To delete the current project tree and the files that are stored in a project, select the project and select **Delete . . .** from the pop-up menu. You can also delete any project by selecting **File** → **Projects** → **Delete . . .**

---

## Opening Existing Projects

To see all of the projects that you have created, select **File** → **Projects** → **Open . . .** Select a project from the list and click **OK** to open it.

---

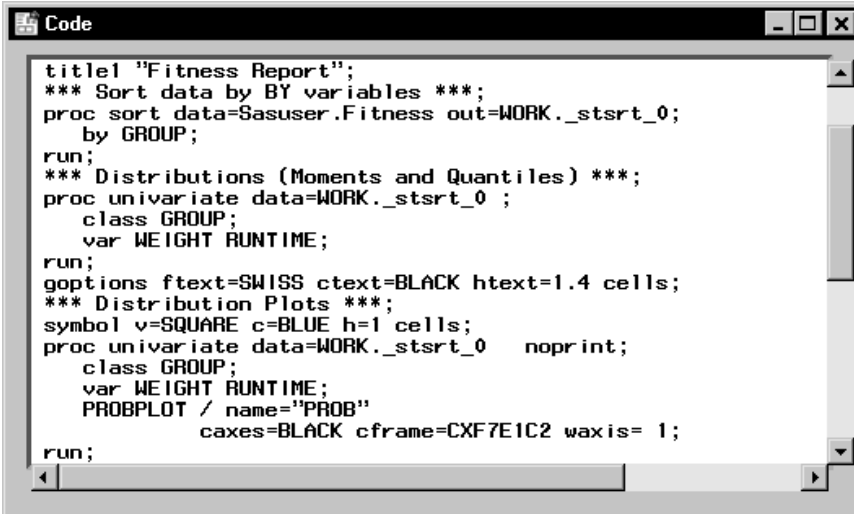
## Using Code

When you perform an analysis or create a graph in Analyst, the code that generated your results is saved in a **Code** node in the project tree. You can view, modify, and submit this code.

---

## Viewing Code in the Code Window

To view the code that generated your results, double-click on a **Code** node in your project tree. The code is displayed in the Code window.



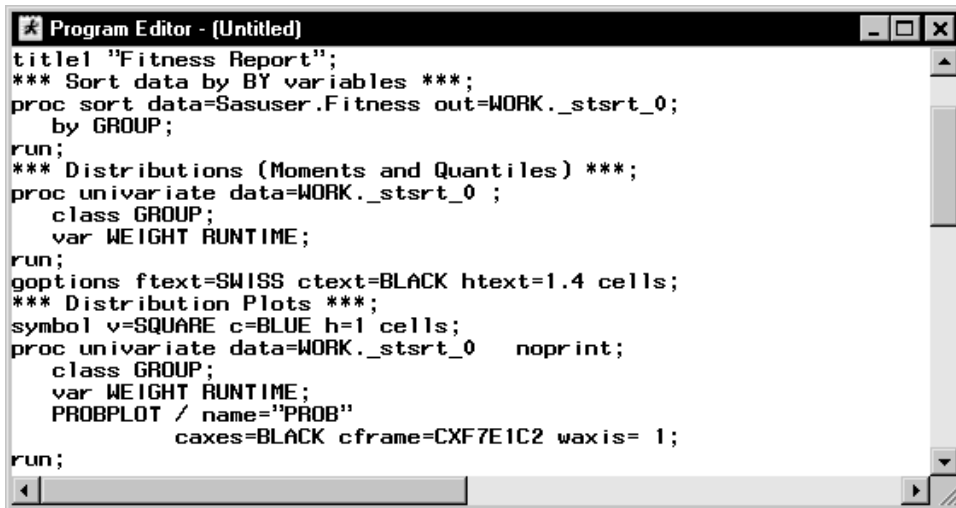
```
Code
title1 "Fitness Report";
*** Sort data by BY variables ***;
proc sort data=Sasuser.Fitness out=WORK._stsrt_0;
  by GROUP;
run;
*** Distributions (Moments and Quantiles) ***;
proc univariate data=WORK._stsrt_0 ;
  class GROUP;
  var WEIGHT RUNTIME;
run;
goptions ftext=SWISS ctext=BLACK htext=1.4 cells;
*** Distribution Plots ***;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=WORK._stsrt_0 noprint;
  class GROUP;
  var WEIGHT RUNTIME;
  PROBPLOT / name="PROB"
           caxes=BLACK cframe=CXF7E1C2 waxis= 1;
run;
```

Figure 3.4. Code Window

---

## Copying Code to the Program Editor Window

To copy code to the Program Editor window, select **Edit** → **Copy to Program Editor** from the Code window.



```

Program Editor - (Untitled)
title1 "Fitness Report";
*** Sort data by BY variables ***;
proc sort data=Sasuser.Fitness out=WORK._stsr0;
  by GROUP;
run;
*** Distributions (Moments and Quantiles) ***;
proc univariate data=WORK._stsr0 ;
  class GROUP;
  var WEIGHT RUNTIME;
run;
goptions ftext=SWISS ctext=BLACK htext=1.4 cells;
*** Distribution Plots ***;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=WORK._stsr0 noprint;
  class GROUP;
  var WEIGHT RUNTIME;
  PROBLOT / name='PROB'
           caxes=BLACK cframe=CXF7E1C2 waxis= 1;
run;

```

**Figure 3.5.** Code in Program Editor Window

In the Program Editor window, you can edit, submit, and save code. Your data must be in browse mode in order for you to submit code that uses the current data table. In edit mode, the data table is locked by Analyst.

---

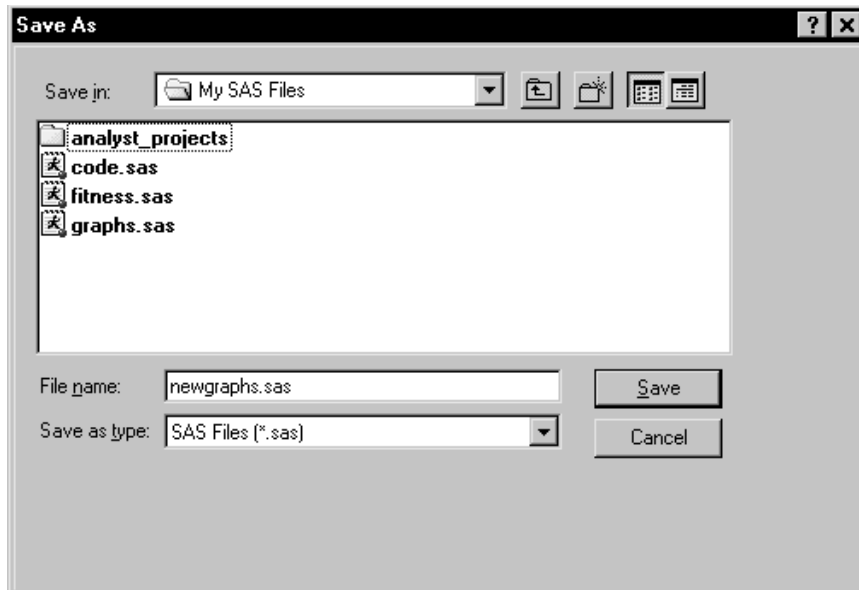
## Printing and Saving Results

You can print and save individual nodes in the project tree.

---

### Saving Text Results

To save code or an analysis result as a file, double-click on a node to open it, and select **File** → **Save As . . .**



**Figure 3.6.** Saving a Text File

Type a filename in the **File name:** field, and select a file type. You can also save code or analysis results by selecting a node and selecting **Save as . . .** from the pop-up menu.

---

## Saving a Graph Result as a File

To save a graph result as a file, double-click on a graph node to open it, and select **File** → **Save As . . .**



**Figure 3.7.** Saving a Graphics File

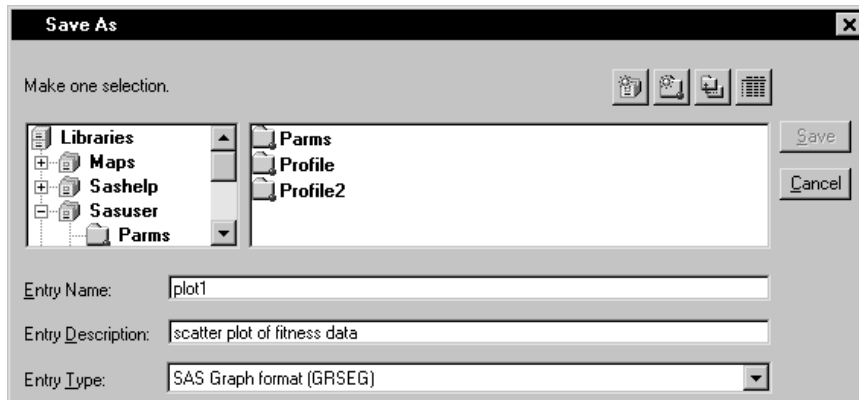
Type a filename in the **File name:** field, and select a file type. You can save the graph in formats that include GIF and postscript.

You can also save a graph result by selecting a node and selecting **Save as . . .** from the pop-up menu.

---

## Saving a Result as a Catalog Entry

To save a result as an entry in a SAS catalog, double-click on the node to open it, and select **File** → **Save as Object . . .**



**Figure 3.8.** Saving a Catalog Entry

Select a library from the list of **Libraries**, and select a catalog. Select an entry name or enter one in the field labeled **Entry Name:**. You can also enter a description for the catalog entry.

---

## Printing Results

You can print code, analysis results, and graph results. Print graph results by opening the graph and selecting **File** → **Print** . . .

To print a code or analysis result, open the node and select **File** → **Print** . . .

---

## Example: Create and Export Histograms

In this example, you open the project that contains the simple regression that you performed in the example at the end of Chapter 1, “Overview,” and save the project under another name. Then you add to the new project by generating histograms from the **Fitness** data.

---

## Open the Project

To open the project that you created in Chapter 1, follow these steps:

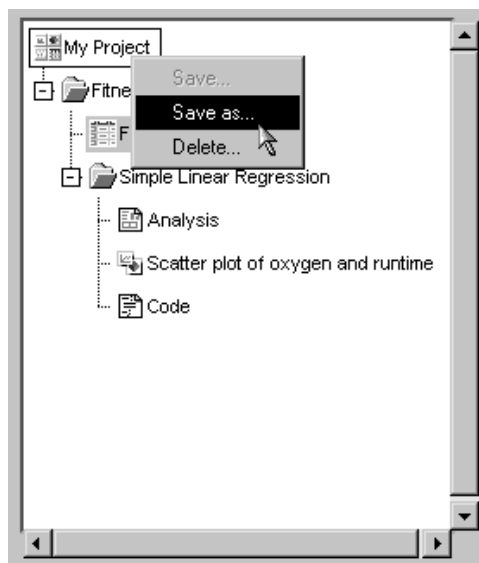
1. Select **File** → **Projects** → **Open . . .**
2. Select **My Project**. Click **OK**.

---

## Save the Project Under Another Name

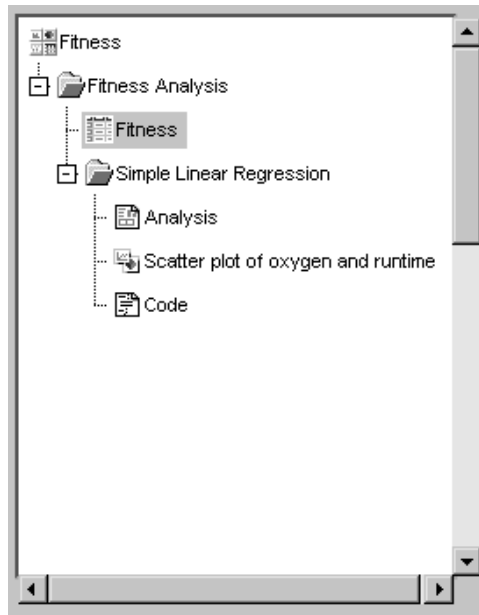
To give the project a more appropriate name, follow these steps:

1. Select **My Project** at the top of the project tree, and select **Save as . . .** from the pop-up menu.



**Figure 3.9.** Saving a Project Under Another Name

2. Type **Fitness** in the **Name:** field and click **OK**.



**Figure 3.10.** Fitness Project

A copy of the project tree is saved with the name **Fitness**. The original project is saved until you delete it.

---

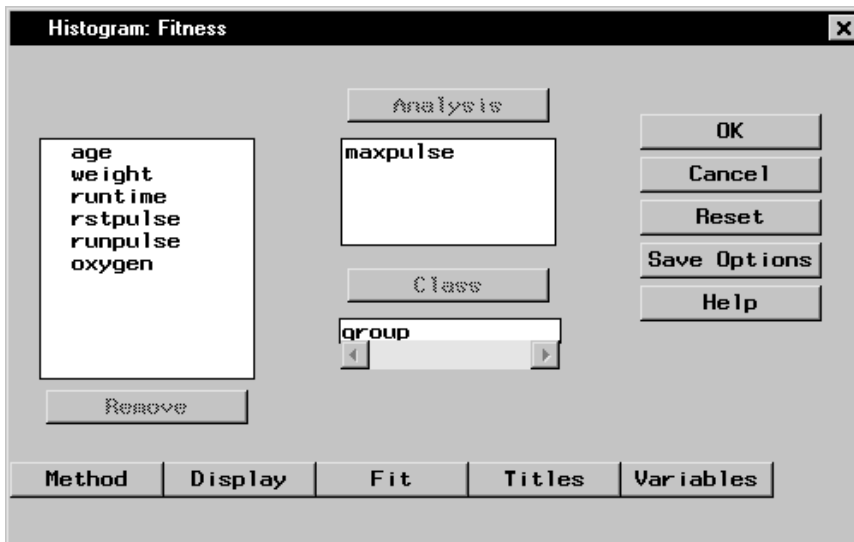
## Generate Histograms

Histograms display the distribution of a particular variable over various intervals, or classes. You can use histograms to see the shape of the distribution and to determine whether the data are distributed symmetrically. A comparative histogram is produced if you specify a classification variable.

To generate comparative histograms of maximum heart rate for each experimental group from the **Fitness** data table, follow these steps:

1. Select **Graphs** → **Histogram ...**
2. Select **maxpulse** from the list, and click on the **Analysis** button. Select **group** from the list, and click on the **Class** button.





**Figure 3.11.** Fitness Analysis and Class Variables

3. To change the way the histogram is displayed, click on the **Display** button.

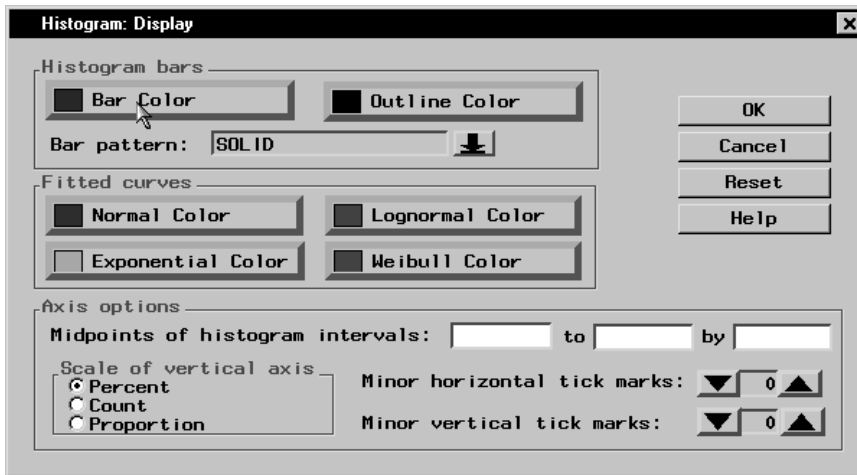
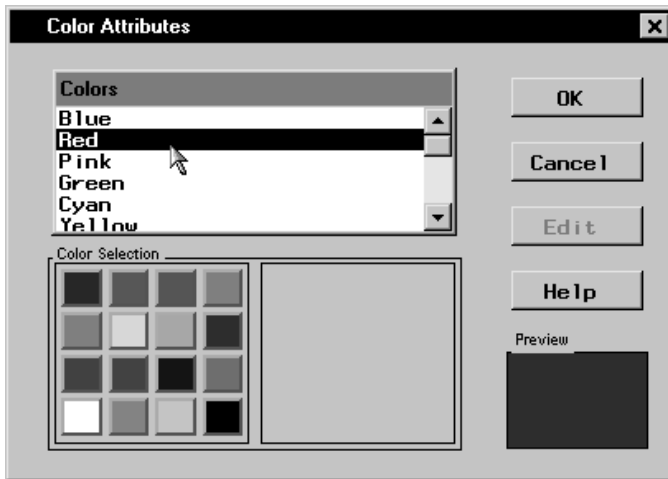


Figure 3.12. Histogram: Display Dialog

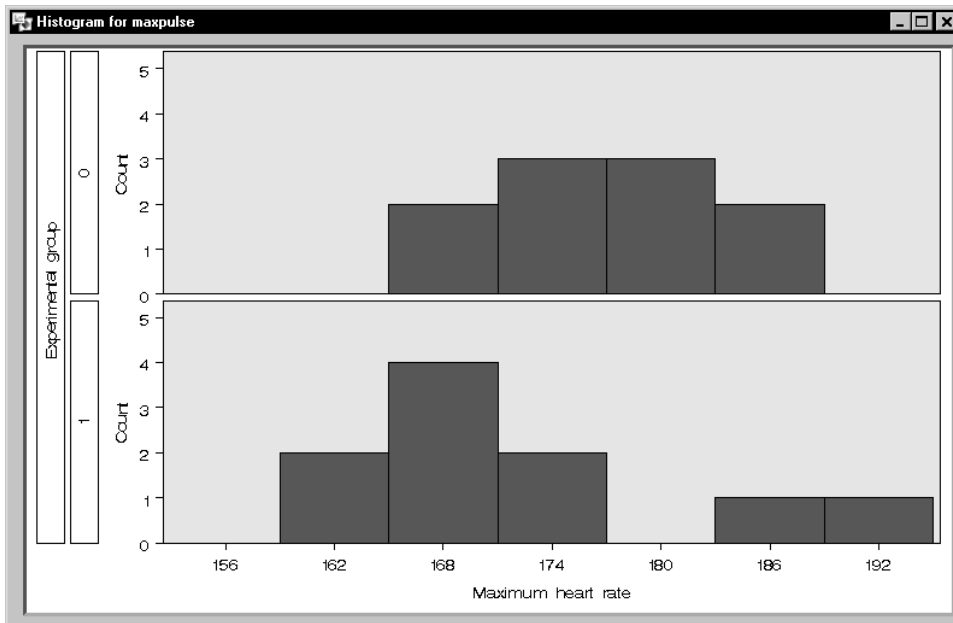
4. Click **Bar Color** to change the color of the histogram bars. Select **Red** from the list of colors.



**Figure 3.13.** Color Attributes Dialog

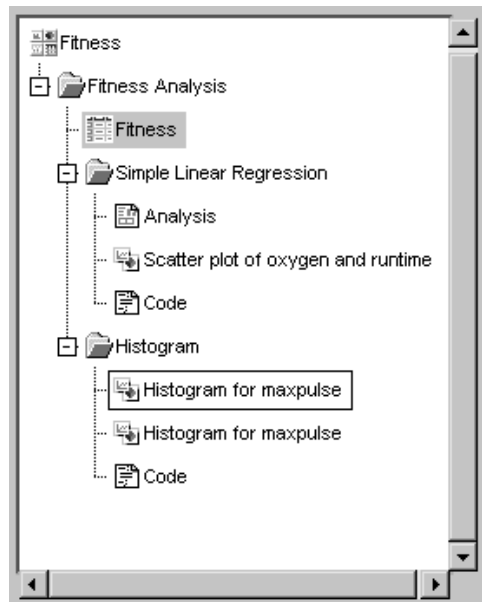
Click **OK** to change the bar color to red.

5. To use number of subjects, rather than percentage, as a gauge of bar size, select **Count** under **Scale of vertical axis**. Click **OK** to return to the Histogram dialog.
6. Click **OK** to create histograms of the maximum heart rate for each group.



**Figure 3.14.** Maximum Heart Rate Histograms

The histograms and the code that produced them have been added to the project tree.



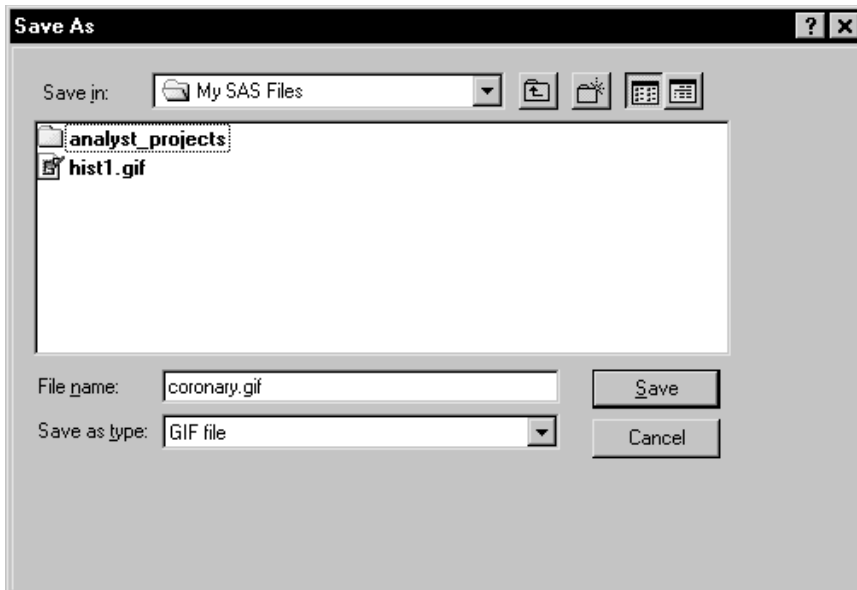
**Figure 3.15.** Project Tree with Histogram Folder

---

## Export Histograms

To save the histogram that you have generated as a graphics file, follow these steps:

1. Double-click on the first node that is labeled **Histogram for maxpulse** to open it.
2. Select **File** → **Save As . . .**
3. In the Save As dialog, click on the arrow next to **Save as type:** and select **GIF file**.
4. Type **coronary.gif** in the **File name:** field.



**Figure 3.16.** Save GIF File

5. Click on the **Save** button to save the file. The histogram is exported to a GIF formatted file.

# Chapter 4

## Customizing Your Session

### Chapter Contents

---

<b>Introduction</b> . . . . .	89
<b>Setting Viewer Preferences</b> . . . . .	89
Window Layout . . . . .	89
Table Settings . . . . .	90
Settings for Variables . . . . .	92
Output Settings . . . . .	93
<b>Setting Graph Preferences</b> . . . . .	94
Point Display Options . . . . .	94
Bar and Contour Rectangle Options . . . . .	96
Axis Options . . . . .	97
Text Options . . . . .	97
<b>Saving Options</b> . . . . .	98
<b>Changing Titles</b> . . . . .	98
<b>Example: Change Global and Task Options</b> . . . . .	100
Change Viewer Settings . . . . .	100
Change Graph Settings . . . . .	105
Change Titles . . . . .	107





# Chapter 4

## Customizing Your Session

---

### Introduction

You can customize your Analyst session from the **Tools** menu by selecting **Viewer Settings . . .** to set viewer preferences and **Graph Settings . . .** to set graph preferences. Any global options that you set are overridden by any individual settings that you specify in a task. These options are also overridden by options that are saved by the **Save Options** button for a task.

You can customize the toolbar by adding other Analyst tasks and icons. See [Chapter 17, “Details,”](#) for information about customizing the Analyst toolbar.

---

### Setting Viewer Preferences

Select **Tools** → **Viewer Settings . . .** to display the Viewer Settings dialog. The Viewer Settings dialog enables you to specify options for the window layout, the data table, and the display of variables and output. When you click **OK**, your changes take effect immediately.

---

### Window Layout

In the **Viewer** tab, you can control the relative size of the project tree and data table by moving the slider at the bottom of the **Window layout** screen.

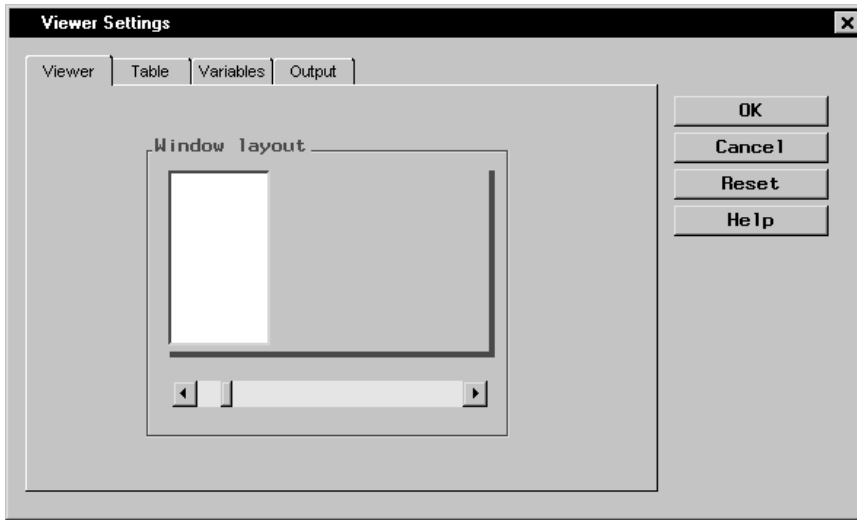


Figure 4.1. Viewer Settings Dialog, Viewer Tab

---

## Table Settings

In the **Table** tab, you can specify the fonts and initial edit mode of your data tables.



**Figure 4.2.** Viewer Settings Dialog, Table Tab

Under **Table fonts**, click on the arrows next to the **Data:** and **Label:** fields to select a font for the data and column headings in the data table.

Under **Show columns with**, select column **Names** or **Labels** to be displayed as column headings.

Under **Open data files for**, specify the mode in which data tables are to be opened. Browse mode prevents any editing of the table. Edit mode allows table editing, and Shared Edit allows multiple users to edit table values concurrently for tables that are accessed through a SAS/SHARE server. These modes can also be changed from the **Edit** menu when the data table is open.

Under **When editing large data files**, you can control processing speed by setting a warning for files that are greater than a certain size.

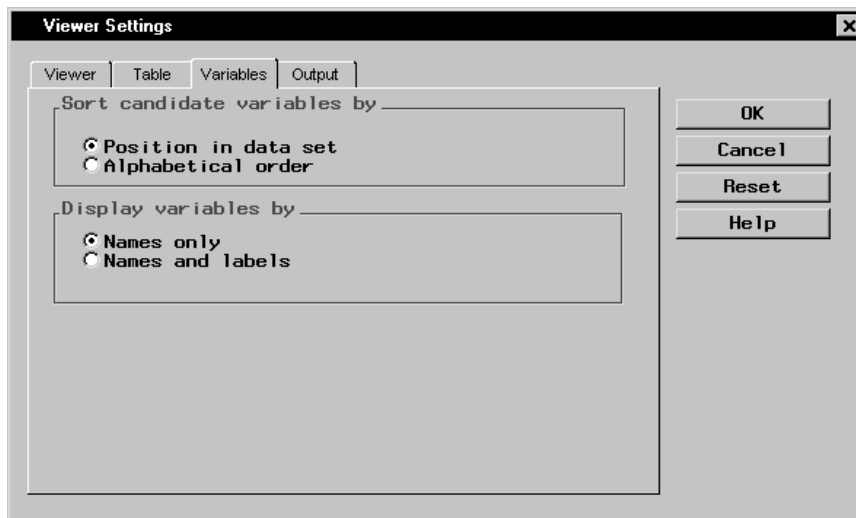
If you have checked **Warn before opening large files to edit**, and the file is larger than the limit you have specified, a message warns you that the data file is large and prompts you to either open a copy or open the data file directly. Opening a copy of the data file takes longer. Opening the data file directly is faster, but changes to the data table cannot be undone.

Click on the up or down arrows to specify the file size limit.

---

## Settings for Variables

In the **Variables** tab, you can customize the display of the variables in the task dialogs.



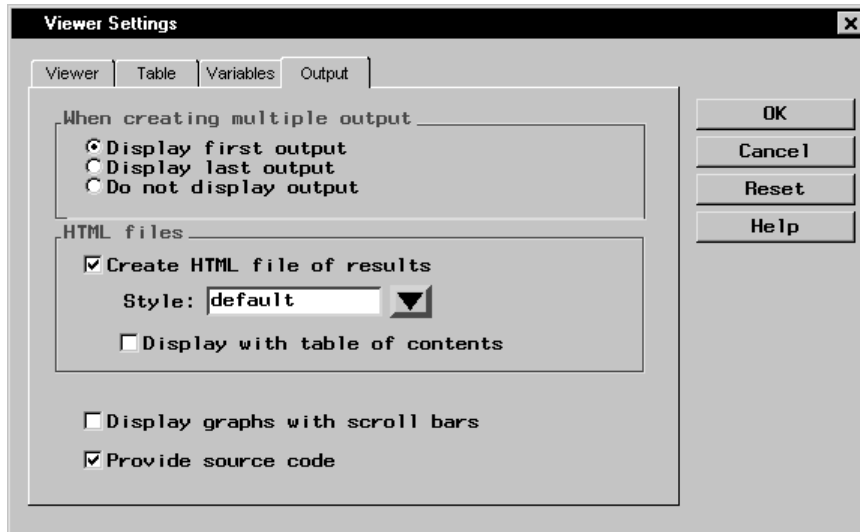
**Figure 4.3.** Viewer Settings Dialog, Variables Tab

Under **Sort candidate variables by**, select **Position in data set** or **Alphabetical order** to specify the order in which to list variable names in the task dialogs.

Under **Display variables by**, select **Names only** or **Names and labels** to specify how variables should be displayed in the task dialogs.

## Output Settings

In the **Output** tab, you can specify options for multiple output, graphs, source code, and HTML files.



**Figure 4.4.** Viewer Settings Dialog, Output Tab

Under **When creating multiple output**, you can determine whether the first or last output should be displayed automatically when an analysis has been run, or whether output should be displayed at all.

Under **HTML files**, select **Create HTML file of results** to include an HTML output node in your project tree whenever you apply a task to your data. You can change the style of the HTML output by selecting a style from the **Style:** drop-down menu. Select **Display with table of contents** to view the HTML output using a table of links to your output (displayed with HTML frames). If this option is not selected, all results are displayed in a single page.

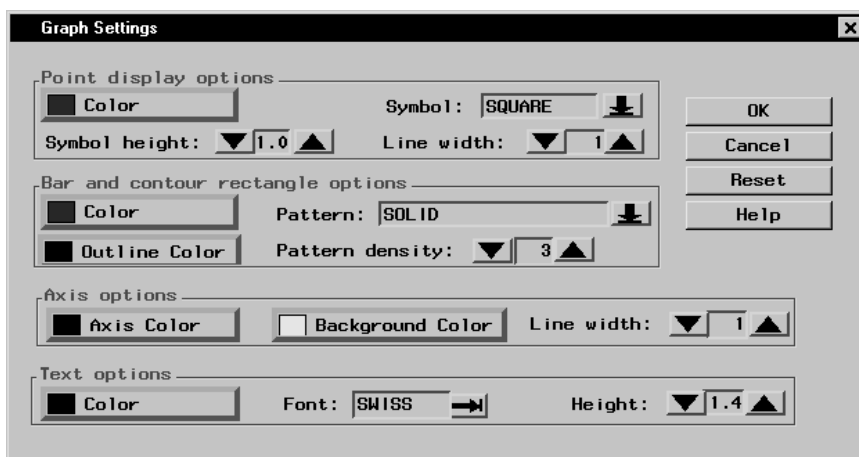
Select **Display graphs with scroll bars** to display scroll bars with your graphs. When scrollbars are displayed, graphs are shown in their natural size. When scrollbars are turned off, graphs are shown in full size in the Output window. Scrollbars can also be turned on or off in the Output window.

Select **Provide source code** to include a source code node in your project tree whenever you apply a task to your data.

---

## Setting Graph Preferences

Select **Tools** → **Graph Settings . . .** to display the Graph Settings dialog. You can use the Graph Settings dialog to customize the appearance of the graphs you produce.



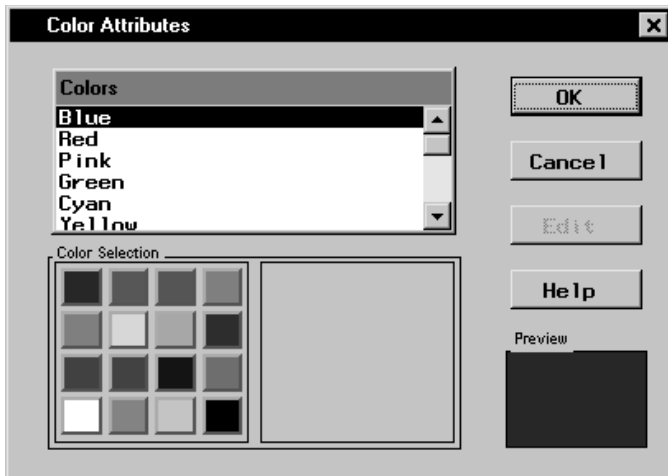
**Figure 4.5.** Graph Settings Dialog

---

### Point Display Options

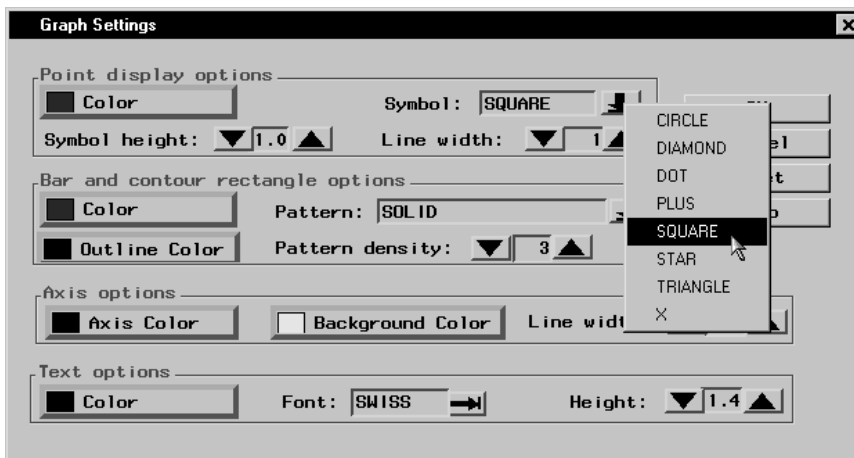
Point display options control the display of points and lines in plots. You can select the color, symbol type, and symbol height of points displayed in the plot. You can also control the color and width of lines in the plot.

Click on the **Color** button to change the color selected to display points.



**Figure 4.6.** Color Attributes Dialog

Click on the arrow next to **Symbol:** to select the symbol used to display points.



**Figure 4.7.** Point Symbols

Click on the down or up arrow next to **Symbol height:** to change the size of the symbol.

Click on the down or up arrow next to **Line width:** to change the width of lines displayed in the plot.

## Bar and Contour Rectangle Options

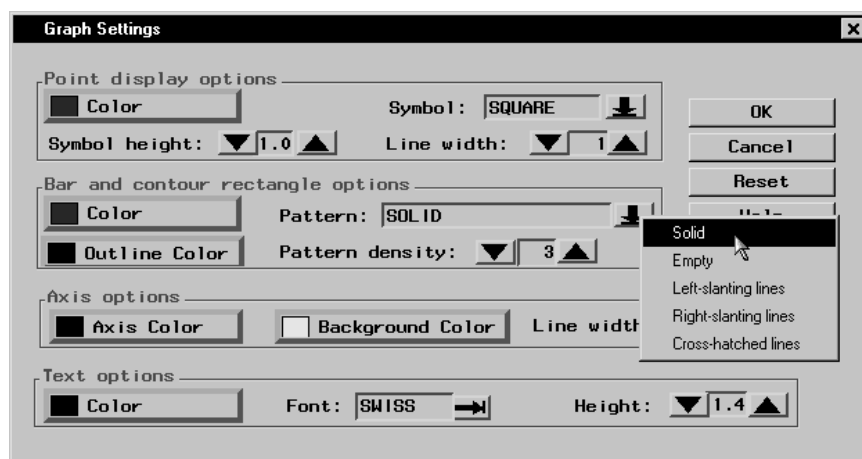
Bar and contour rectangle options control the display of any bars or rectangles in graphs. You can control the outline color, the fill color, the fill pattern, and the pattern density.

Click on the **Color** button to select the color used to fill bars and rectangles.

Click on the **Outline Color** button to select the color used for bar outlines.

Click on the down arrow next to **Pattern:** to select a pattern used to fill bars and rectangles.

Click on the down or up arrow next to **Pattern density:** to change the density of the pattern.



**Figure 4.8.** Pattern Choices



## Axis Options

Axis options control the color and width of axis lines as well as the background color of the graph.

Click on the **Axis Color** button to select the color used for axis lines.

Click on the **Background Color** button to change the background color of the graph.

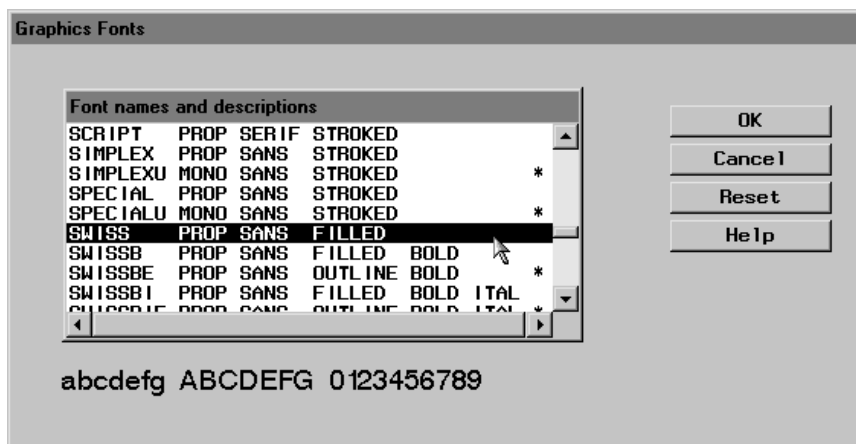
Click on the down or up arrow next to **Line width:** to change the width of the axis lines.

## Text Options

Text options control the color, font, and size of any text in the graph.

Click on the **Color** button to change the color used for text.

Click on the arrow next to **Font:** to select a text font. Do not pick a font for which no sample text is displayed.



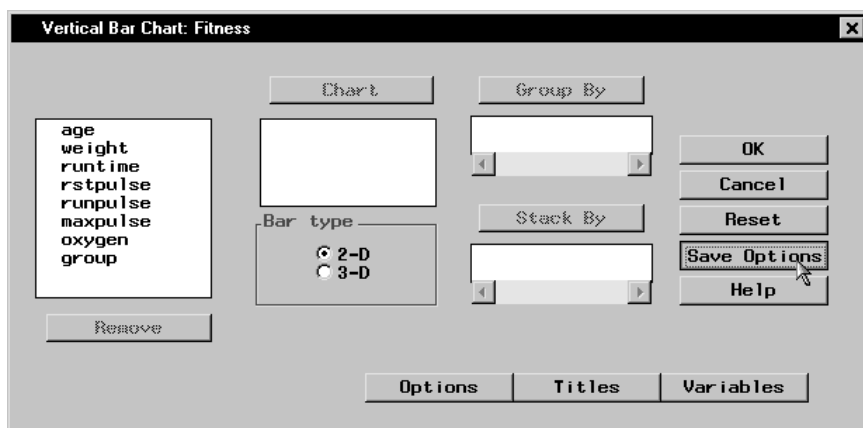
**Figure 4.9.** Graphics Fonts Dialog

Click on the down or up arrow next to **Height:** to change the text height.

---

## Saving Options

You can save any option that is associated with a task by clicking on the **Save Options** button in the task dialog. For example, you can save the options that are associated with the Bar Chart task by clicking on the **Save Options** button in the Bar Chart dialog.



**Figure 4.10.** Save Options Button

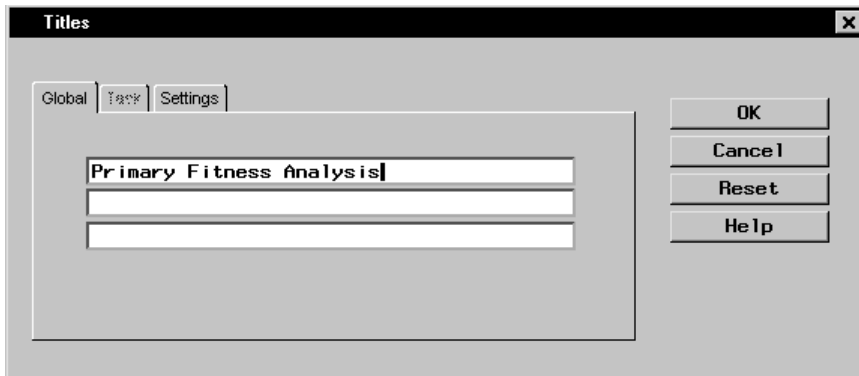
These options become your defaults and are applied when you click on the **Reset** button. These options are also saved between sessions.

Options that are associated with data, such as **Group By** variables, cannot be saved with the task options, and do not persist between sessions.

---

## Changing Titles

Select **Tools** → **Titles . . .** or click on the **Titles** button within a task to specify the titles that appear on the output.



**Figure 4.11.** Titles Dialog, Global Tab

In the **Global** tab, you can specify titles that are displayed on all output. These titles are saved across Analyst sessions.

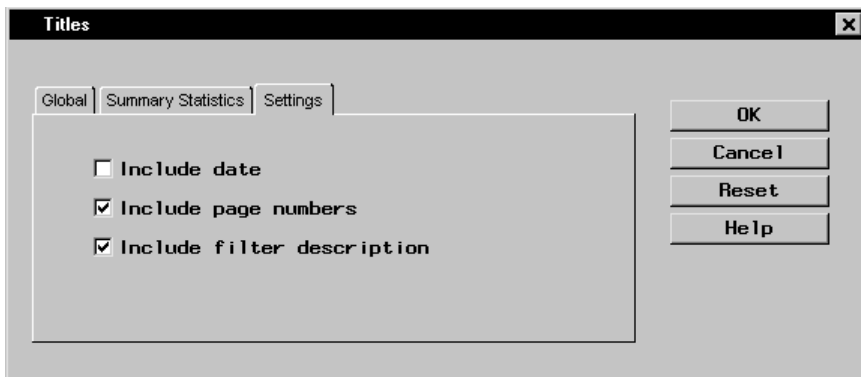
If you have selected the **Titles** button within a task, you can use the tab for the current task to specify titles for the output from the task. For example, if you are in the Summary Statistics task, you can specify the titles for the output from that task.



**Figure 4.12.** Titles Dialog, Task Tab

Select the box next to **Override global titles** to exclude the global titles from the task results.

In the **Settings** tab, you can specify whether or not to include the date, the page numbers, and a filter description.



**Figure 4.13.** Titles Dialog, Settings Tab

Global titles information and settings are saved between SAS sessions.

---

## Example: Change Global and Task Options

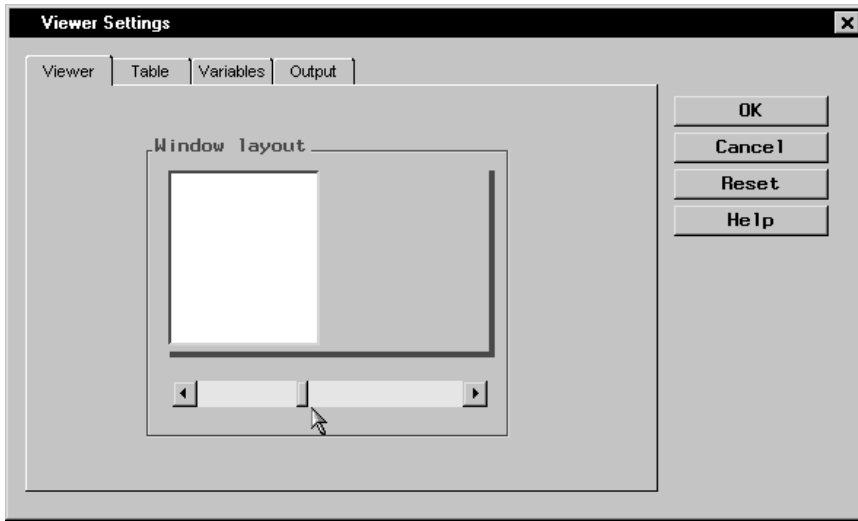
In this example, you change the viewer and graph settings, and the titles that appear on your output.

---

### Change Viewer Settings

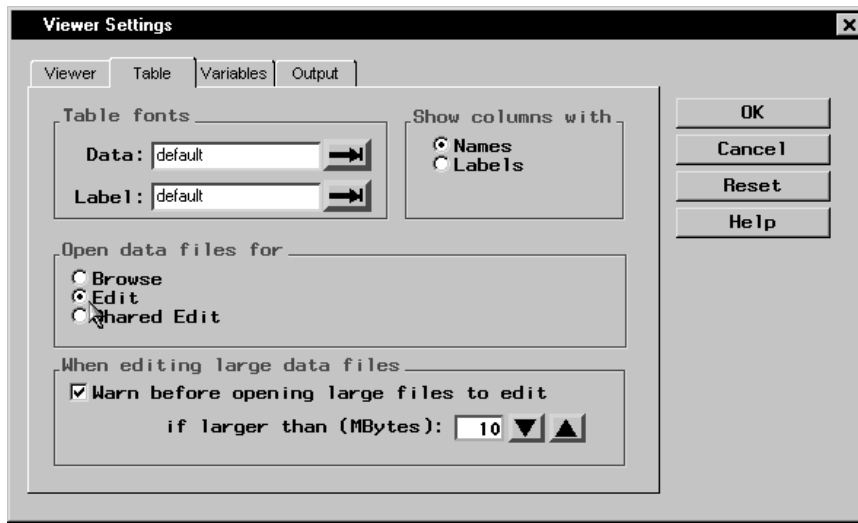
To change the window layout, open data files automatically in edit mode, display candidate variables in alphabetical order, and create HTML files of your results, follow these steps:

1. To change the window layout to make long node names easier to read, select **Tools** → **Viewer Settings** . . . Move the slider to the right so that the project tree is displayed in a wider window.



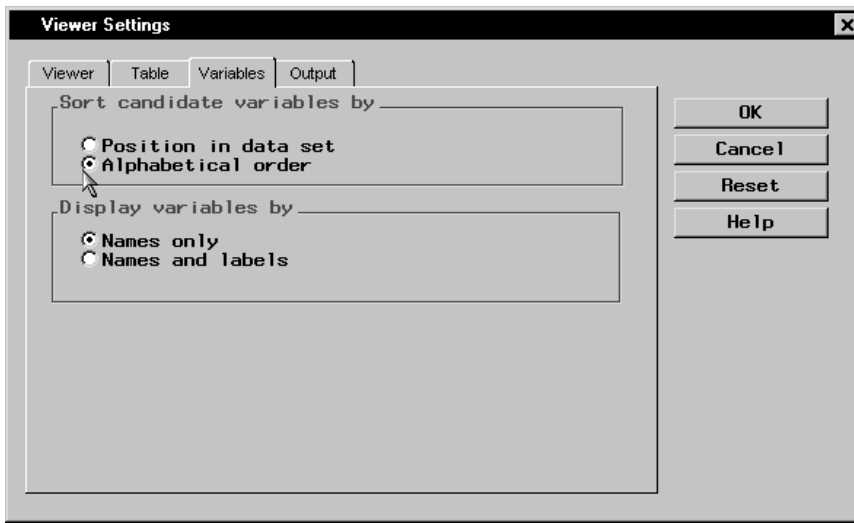
**Figure 4.14.** Wider Project Tree Window Setting

2. To automatically open data files in edit mode, select the **Table** tab, and select **Edit** under the **Open data files for** heading.



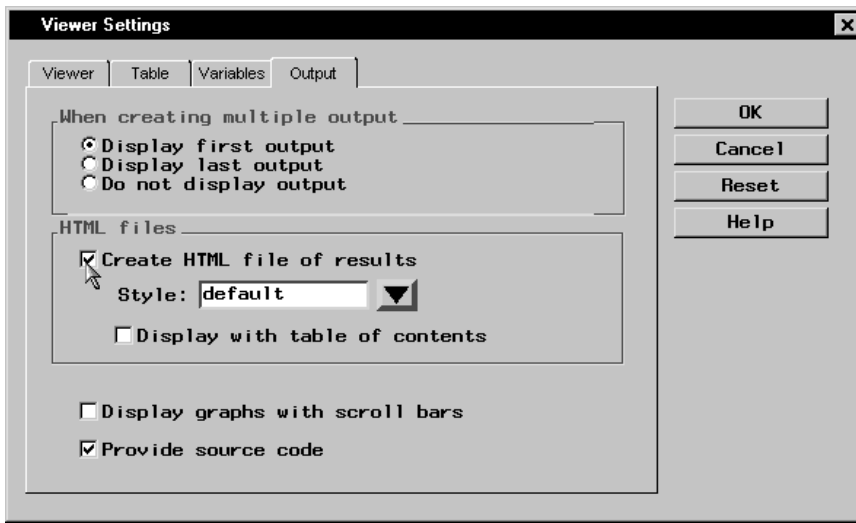
**Figure 4.15.** Open Data Files for Edit

3. To display the candidate variables in alphabetical order in a task dialog, select the **Variables** tab and select **Alphabetical order** under the **Sort candidate variables by** heading.



**Figure 4.16.** Sort Candidate Variables by Alphabetical Order

4. To automatically create HTML files of your results, select the **Output** tab and select **Create HTML file of results** under the **HTML files** heading.

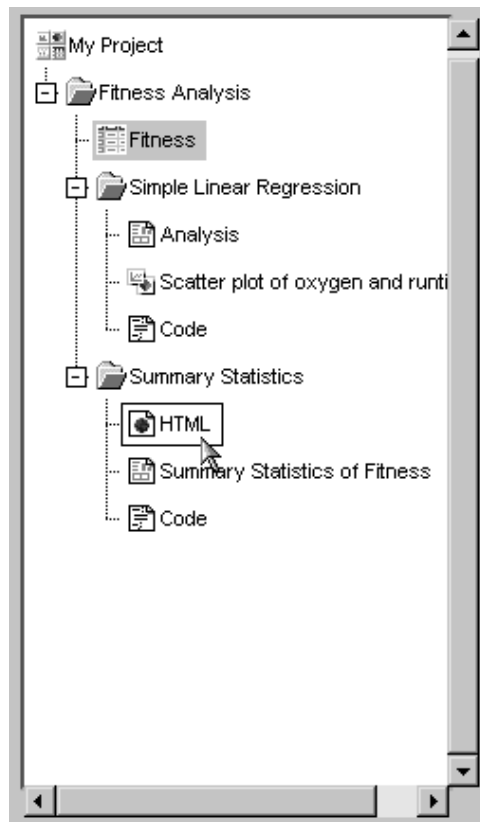


**Figure 4.17.** Create HTML File of Results

5. Click **OK** to save your viewer settings.

When you run an analysis, the HTML results are displayed as a separate node in the project tree.





**Figure 4.18.** HTML Results Node

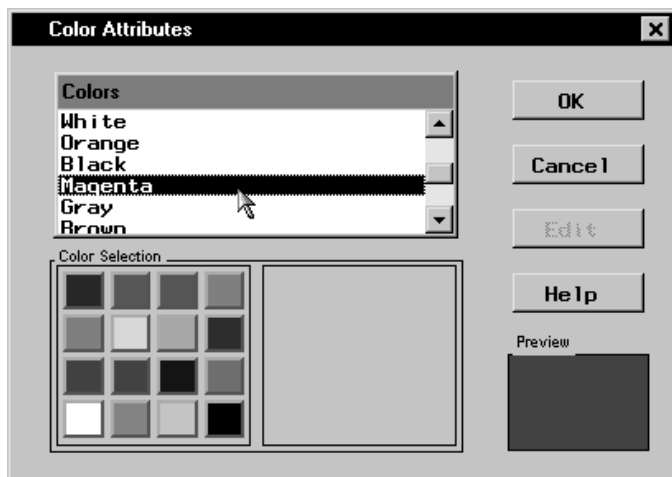
If you double-click on the HTML results, they are displayed in your HTML browser.

---

## Change Graph Settings

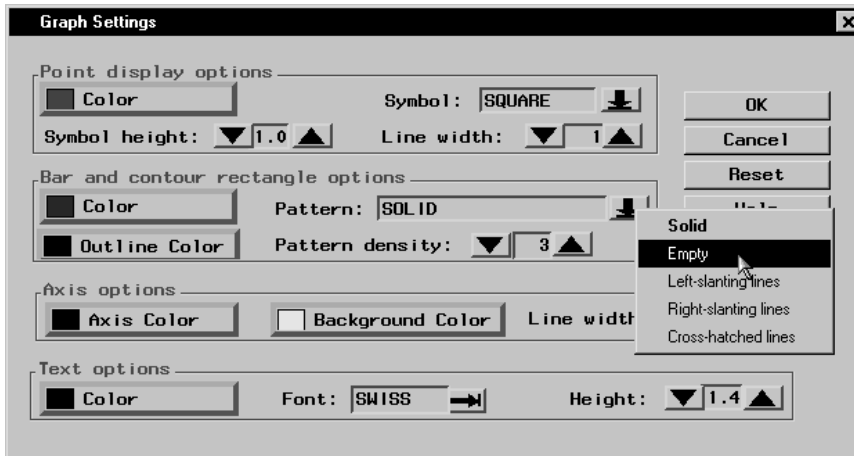
To change the point and line display color and the bar and contour rectangle pattern in your graphs, follow these steps:

1. Select **Tools** → **Graph Settings** ...
2. Select **Color** under **Point display options**.
3. Select **Magenta** from the list of colors.



**Figure 4.19.** Select Point and Line Display Color

- Click **OK**.
4. Under the **Bar and contour rectangle options** heading, click on the arrow next to **Pattern:** and select **Empty** from the list of patterns.



**Figure 4.20.** Select Bar and Contour Rectangle Pattern

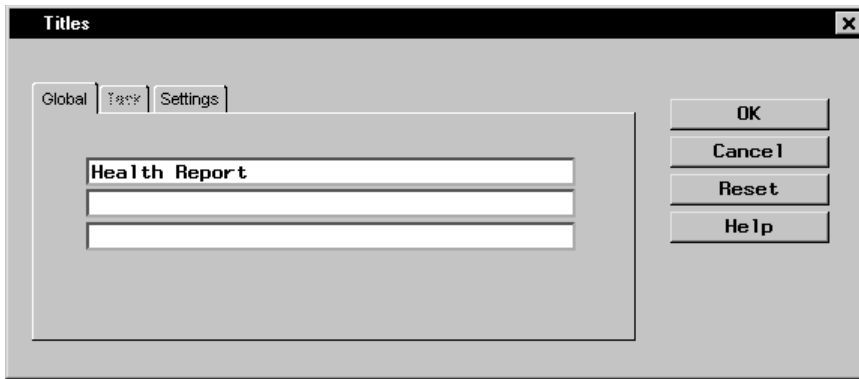
5. Click **OK** to save your graph settings.

---

## Change Titles

To specify a default title for all your output, follow these steps:

1. Select **Tools** → **Titles . . .**
2. Under the **Global** tab, type **Health Report** in the first field.



**Figure 4.21.** Specifying a Global Title

3. Click **OK** to apply this title to all subsequent output.

# Chapter 5

## Creating Graphs

### Chapter Contents

---

<b>Introduction</b> . . . . .	111
<b>Bar Charts</b> . . . . .	111
Bar Chart Options . . . . .	112
Bar Chart Titles . . . . .	117
Bar Chart Variables . . . . .	118
Example: Create a 3-D Bar Chart . . . . .	119
<b>Pie Charts</b> . . . . .	128
Pie Chart Options . . . . .	128
Pie Chart Titles . . . . .	134
Pie Chart Variables . . . . .	134
Example: Create a 3-D Pie Chart . . . . .	134
<b>Scatter Plots</b> . . . . .	141
Two-Dimensional Scatter Plot Options . . . . .	142
Three-Dimensional Scatter Plot Options . . . . .	144
Scatter Plot Titles . . . . .	145
Scatter Plot Variables . . . . .	145
Example: Create a 2-D Scatter Plot . . . . .	146



# Chapter 5

## Creating Graphs

### Introduction

In the Analyst Application, you can use bar charts, pie charts, and scatter plots, in addition to other kinds of graphs, to display your data graphically. Vertical and horizontal bar charts display your data in the form of a two-dimensional or three-dimensional bar graph. A pie chart displays your data in the form of a two-dimensional or three-dimensional disc, divided into slices. The size of each slice indicates the relative contribution of each part to the whole. A scatter plot displays any relationship between two or more variables.

### Bar Charts

To create a bar chart, select **Graphs** → **Bar Chart**. Select **Horizontal . . .** or **Vertical . . .** to create a horizontal or a vertical bar chart.

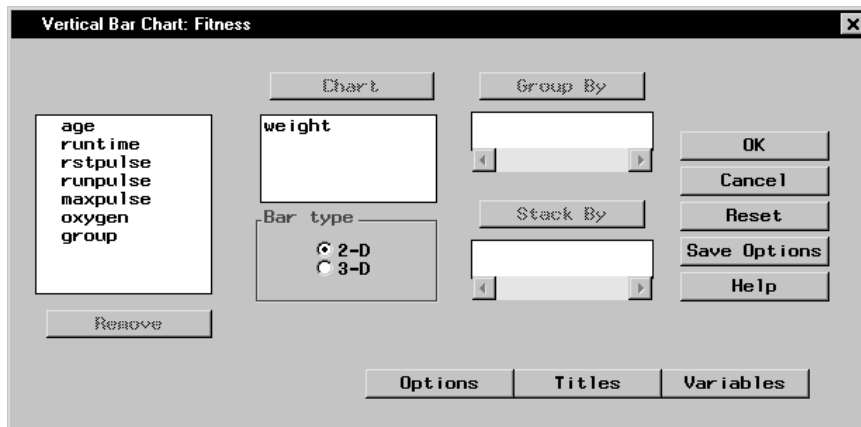


Figure 5.1. Vertical Bar Chart Dialog

Select variables from the candidate list and click on the **Chart** button to create bar charts of those variables.

Select **2-D** or **3-D** under **Bar type** to specify whether you want to display a two-dimensional or a three-dimensional chart.

Select a variable from the candidate list and click on the **Group By** button to add the variable to be used as a grouping variable in the bar chart. This organizes the bars into groups based on the values of the grouping variable.

Select a variable from the candidate list and click on the **Stack By** button to add the variable to be used as a stacking variable in the bar chart. Using a stacking variable subdivides, or stacks segments of, each bar based on the contribution of the stacking variable.

---

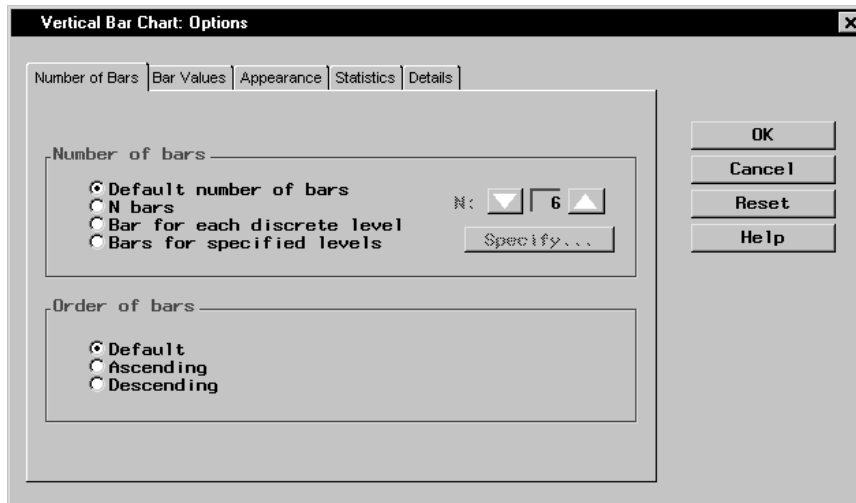
## Bar Chart Options

Click on the **Options** button to display the Bar Chart Options dialog. In the Bar Chart Options dialog, you can control the appearance of your horizontal or vertical bar chart. Click **OK** to save your changes.

### *Number of Bars*

The **Number of Bars** tab enables you to specify the number of bars in the chart and the order in which they are displayed.





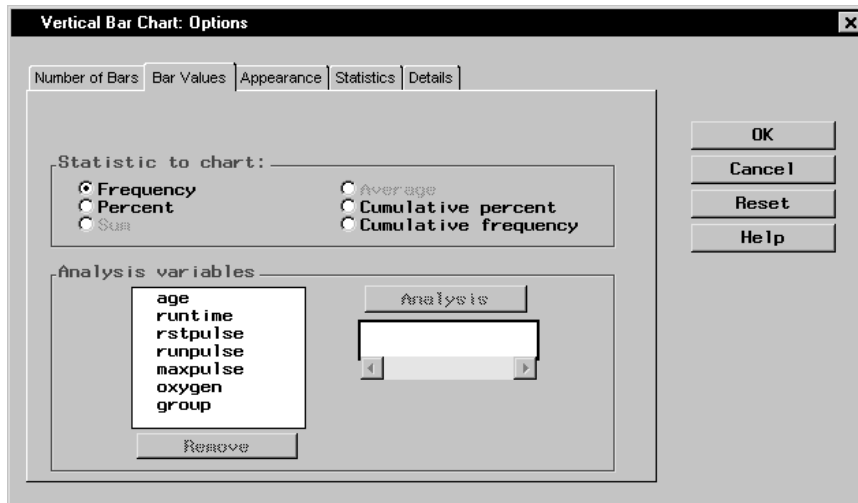
**Figure 5.2.** Number of Bars Tab

Select **Default number of bars** to display a default number of bars based on the chart variable. Select **N bars** and select a number from the list to specify the number of bars to be displayed. Select **Bar for each discrete level** to display a bar for each discrete level of the chart variable. If there is only one chart variable, select **Bars for specified levels** and click on the **Specify** button to provide a list of midpoints or to specify a range of numeric values, or to provide a list of character values.

Under **Order of bars**, select **Default**, **Ascending**, or **Descending** to display your data in default order, ascending order of bar length, or descending order of bar length.

### **Bar Values**

The **Bar Values** tab enables you to control the type of information that is displayed by each bar by specifying the statistic to display in the chart and any additional variable to use in computing the statistic.



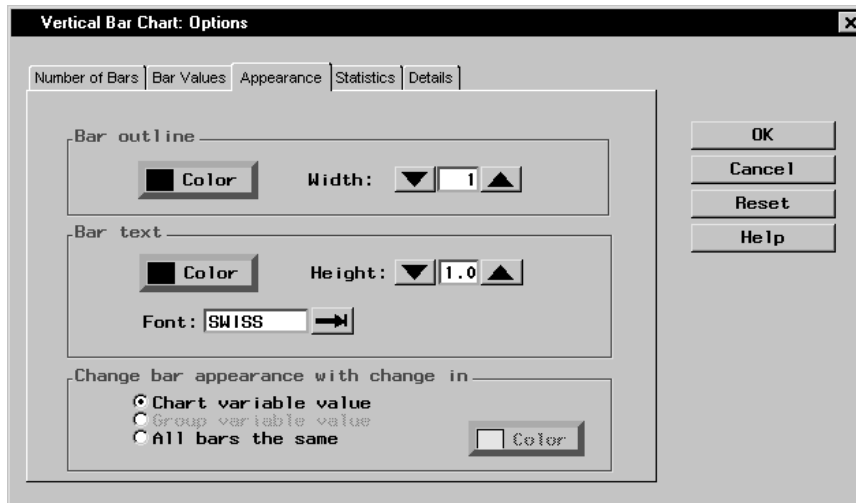
**Figure 5.3.** Bar Values Tab

If you do not specify an analysis variable, you can select frequency, percent, cumulative percent, or cumulative frequency as the statistic to chart. Each bar represents the selected statistic for the current midpoint value of the chart variable.

If you specify an analysis variable, you can select sum or average as the statistic to chart. Each bar displays the sum or average of the analysis variable for the current midpoint value of the chart variable.

### **Appearance**

The **Appearance** tab enables you to select colors and fonts.



**Figure 5.4.** Appearance Tab

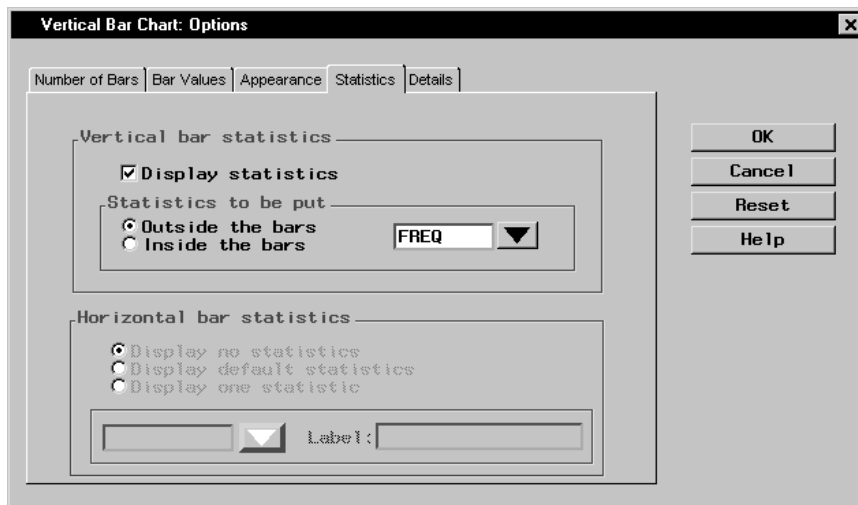
Under **Bar outline**, click on the **Color** button and select a color for the outline of the bar from the Color Attributes dialog. Specify the width of the bar outline in pixels in the **Width:** selector.

Under **Bar text**, click on the **Color** button and select a color for the chart text from the Color Attributes dialog. Specify the height of the text in cells in the **Height:** selector. Select a font by clicking on the arrow next to the **Font:** selector.

Under **Change bar appearance with change in**, you can track changes in the chart or group variable values by color, or you can choose to have all bars remain the same color. If you choose **All bars the same**, you can specify the color to be used.

### **Statistics**

The **Statistics** tab enables you to specify the display of statistics in horizontal and vertical bar charts.



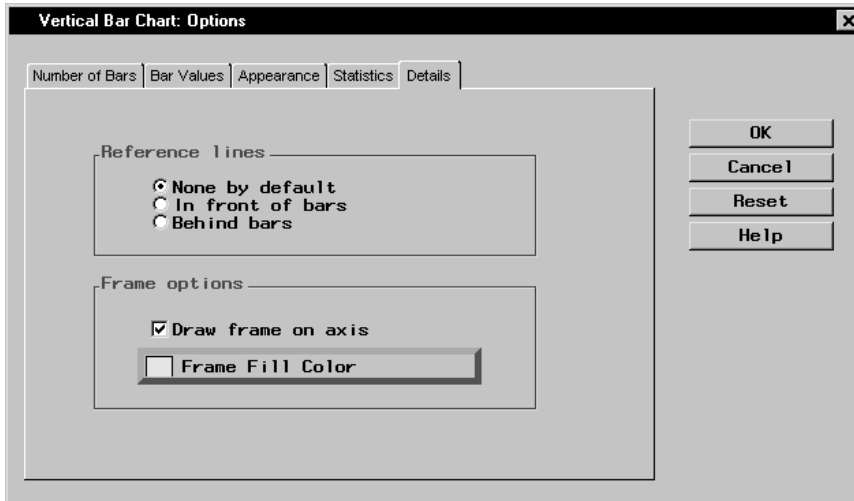
**Figure 5.5.** Statistics Tab

If the chart is a vertical bar chart, the **Vertical bar statistics** section is clickable and the **Horizontal bar statistics** section is greyed. Select **Display statistics** if you want statistics to be displayed in the chart, and specify whether the statistics should be displayed inside or outside the bars of the chart. Select the statistic to be displayed from the list.

If the chart is a horizontal bar chart, the **Horizontal bar statistics** section is clickable and the **Vertical bar statistics** section is greyed. Select **Display no statistics** to hide statistics from display. Select **Display default statistics** to display the statistics that have been applied to the chart. To display one statistic, select **Display one statistic**, and select the statistic to be displayed from the list.

### **Details**

The **Details** tab enables you to specify reference lines and frame options.



**Figure 5.6.** Details Tab

Under **Reference lines**, you can select whether to display no reference lines, or display reference lines in front of or behind the bars in the chart.

Under **Frame options**, when you select **Draw frame on axis**, you can click on the **Frame Fill Color** button and select a color for the frame from the Color Attributes dialog.

---

## Bar Chart Titles

Click on the **Titles** button to display the Titles dialog.



**Figure 5.7.** Titles Dialog, Bar Chart Tab

In the **Global** tab, you can specify titles that are displayed on all output. These titles are saved across Analyst sessions.

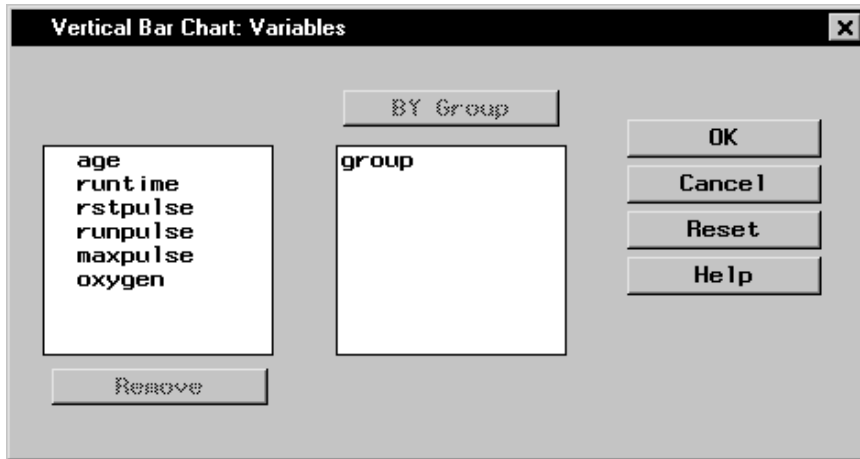
In the **Bar Chart** tab, you can specify titles for the bar chart. Select the box next to **Override global titles** to exclude the global titles from the bar chart results.

In the **Settings** tab, you can specify whether or not to include the date, the page numbers, and a filter description.

---

## Bar Chart Variables

Click on the **Variables** button to display the Bar Chart Variables dialog.



**Figure 5.8.** Vertical Bar Chart: Variables Dialog

BY group variables separate the data set into groups of observations. Separate analyses are performed for each group and displayed in separate charts. For example, you could use a BY group variable to perform separate analyses on females and males. Specify BY group variables by selecting them in the candidate list and clicking on the **BY Group** button.

---

## Example: Create a 3-D Bar Chart

### *Open the Fitness Data Set*

In this example, you create a bar chart using the Fitness data set. To open the Fitness data set, follow these steps:

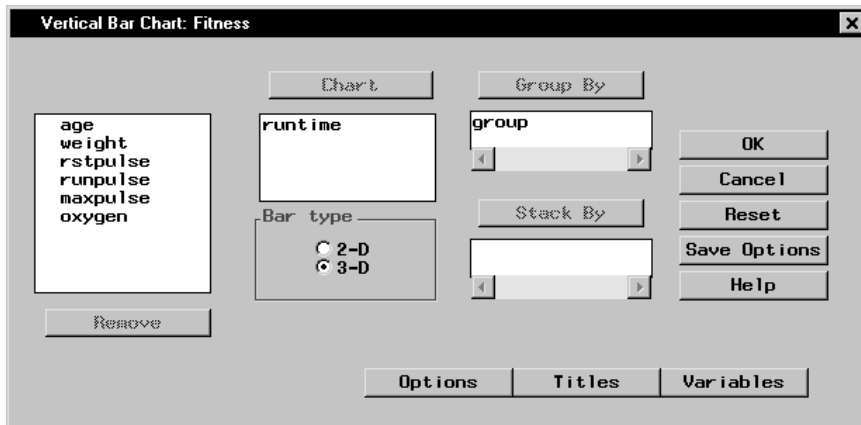
1. Select **Tools** → **Sample Data . . .**
2. Select **Fitness**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Fitness** from the list of members.
7. Click **OK** to bring the **Fitness** data set into the data table.

### **Specify Chart and Grouping Variables**

To create a 3-D vertical bar chart that compares among experimental groups the average amount of oxygen consumed given the time it takes to run 1.5 miles, follow these steps:

1. Select **Graphs** → **Bar Chart** → **Vertical . . .** to display the Vertical Bar Chart dialog.
2. Select **runtime** from the candidate list, and click **Chart** to make minutes to run 1.5 miles the charted variable.
3. Under **Bar type**, select **3-D** to make the bar chart three-dimensional.
4. To compare among experimental groups, select **group** from the candidate list and click **Group By**.



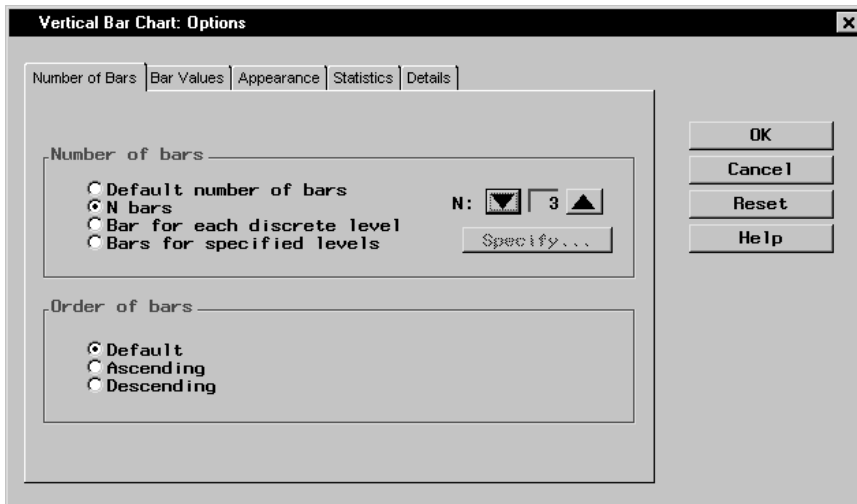


**Figure 5.9.** Chart and Grouping Variables

### **Specify Bar Chart Options**

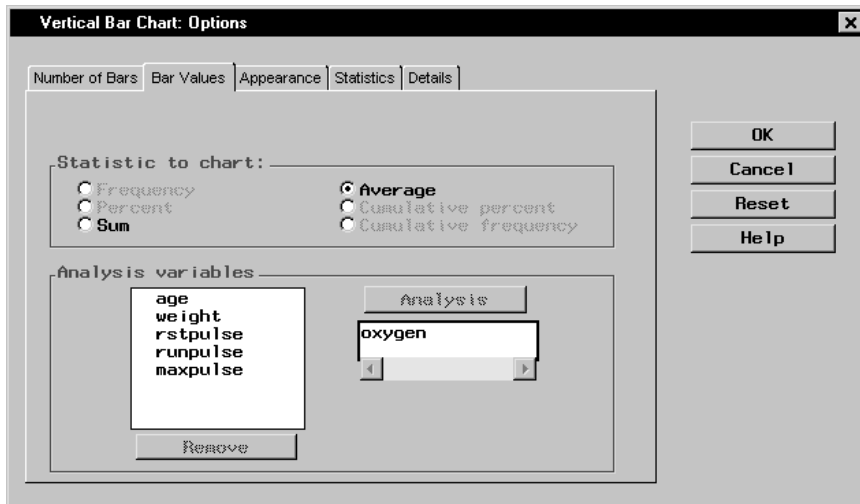
To specify your bar chart options, such as the number and appearance of the bars, follow these steps.

1. Click on the **Options** button to display the Bar Chart Options dialog.
2. Under **Number of bars**, select **N bars**, and click on the down arrow until **N** = 3. Because a grouping variable was specified, bars for three runtime midpoints are displayed for each value of the experimental group.



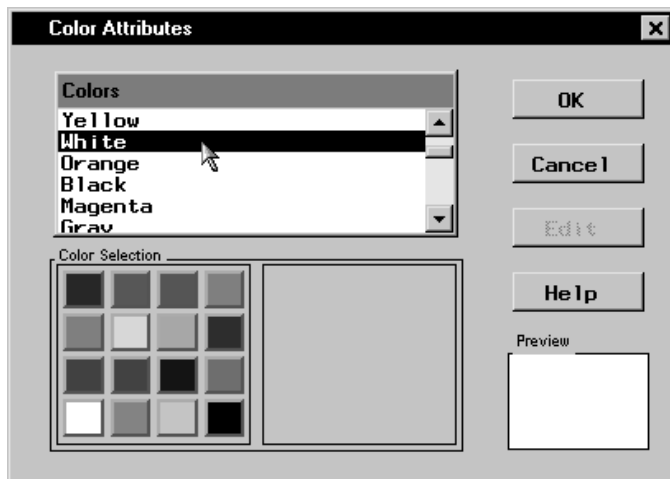
**Figure 5.10.** Number of Bars

3. Select the **Bar Values** tab. Under **Analysis variables**, select **oxygen** from the candidate list and click on the **Analysis** button to make oxygen consumption your analysis variable.
4. Under **Statistic to chart**, select **Average** to display the average oxygen consumption per runtime.



**Figure 5.11.** Bar Values

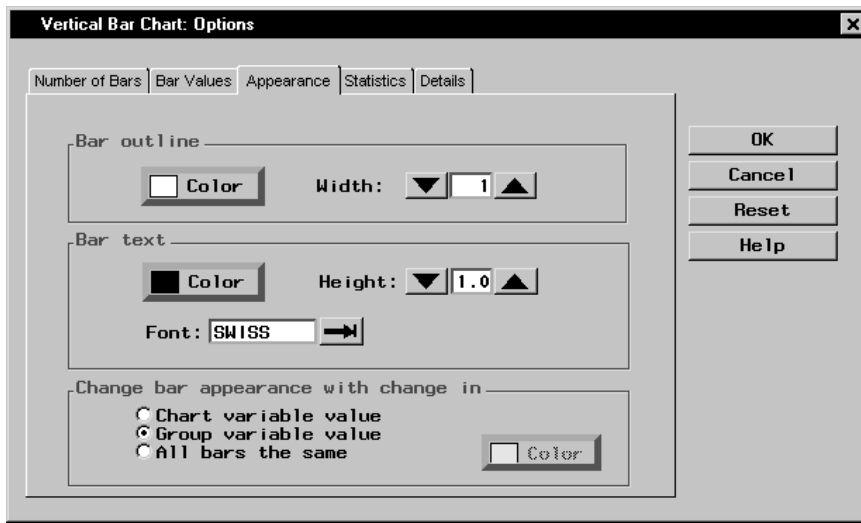
5. Select the **Appearance** tab. Under **Bar outline**, click on the **Color** button. Select **White** from the Color Attributes list to make the bar outlines white.



**Figure 5.12.** Bar Outlines

Click **OK** to close the Color Attributes window and return to the Bar Chart Options dialog.

6. Still on the **Appearance** tab, select **Group variable value** under **Change bar appearance with change in**.



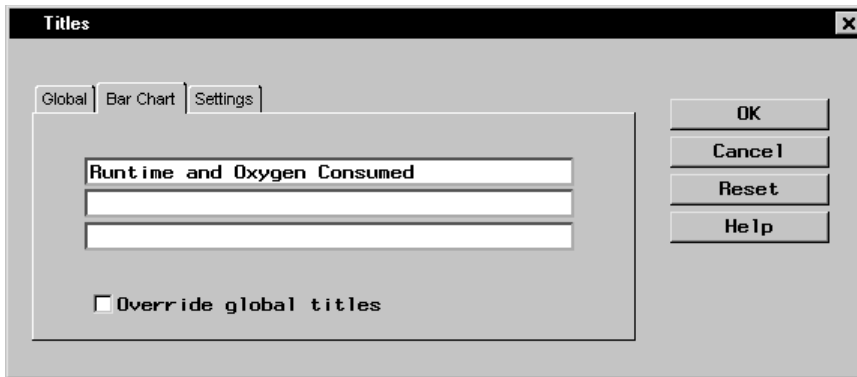
**Figure 5.13.** Bar Appearance

7. Click **OK** to return to the Vertical Bar Chart dialog.

### **Specify Bar Chart Titles**

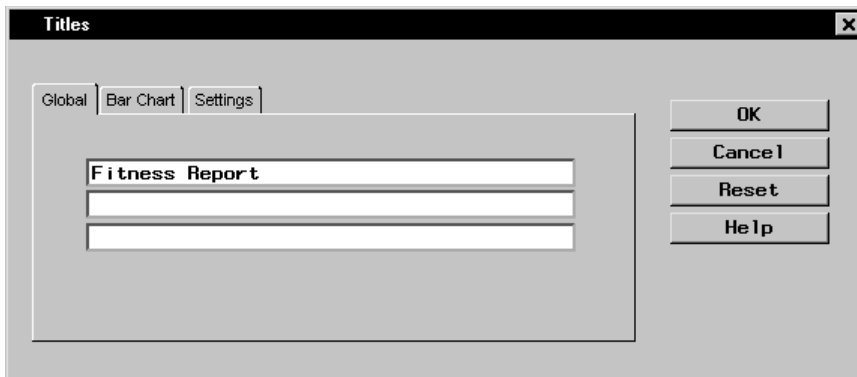
To specify the titles for your bar chart, follow these steps:

1. Click on the **Titles** button in the Vertical Bar Chart dialog.
2. In the **Bar Chart** tab, type **Runtime and Oxygen Consumed** in the first field.



**Figure 5.14.** Bar Chart Title

3. Click on the **Global** tab. Type **Fitness Report** in the first field. This global title is saved across all Analyst sessions until you change it.

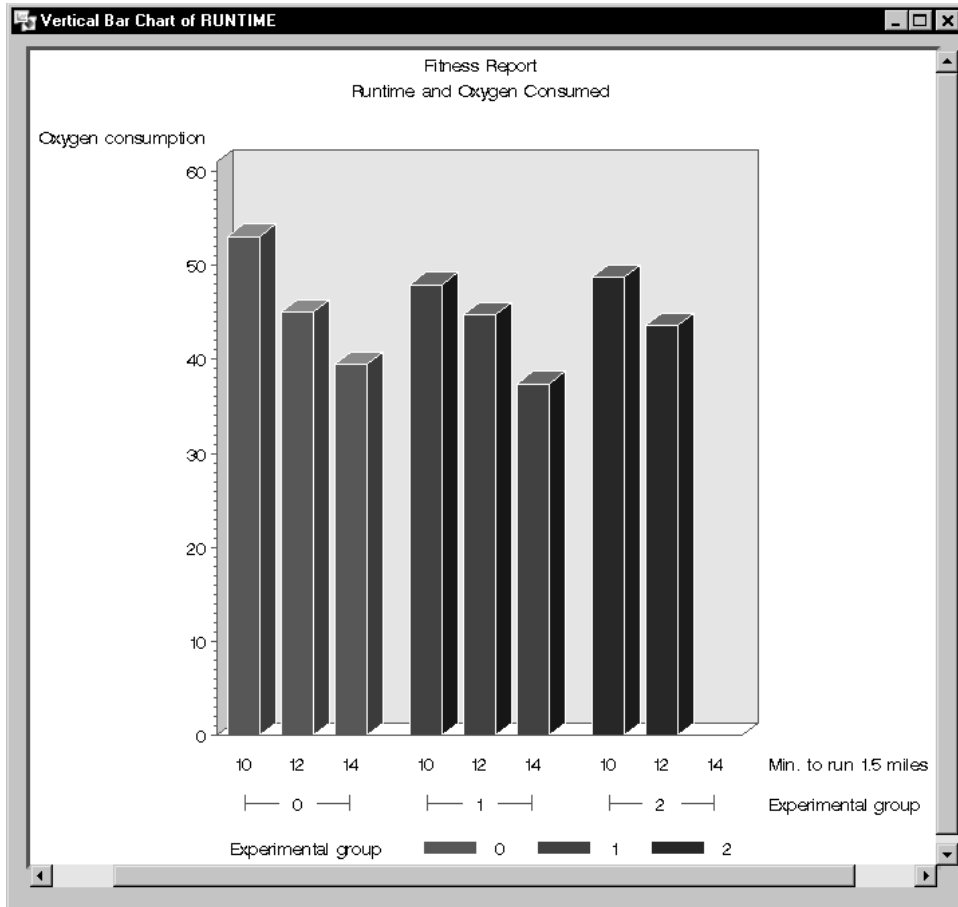


**Figure 5.15.** Global Title

4. Click **OK** to save your title changes.

### **Generate Bar Chart**

To display your bar chart, click **OK** in the Vertical Bar Chart dialog.



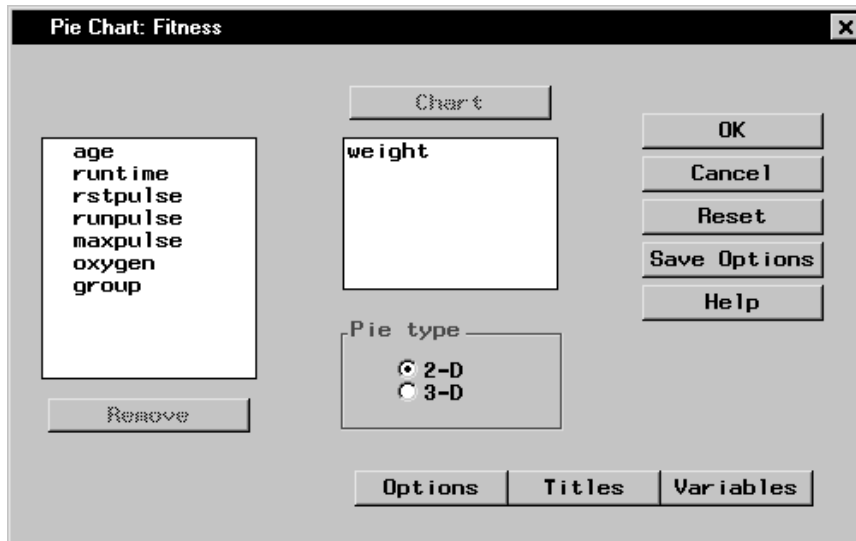
**Figure 5.16.** Vertical Bar Chart

As expected, larger amounts of oxygen are consumed by faster runners. Experimental group does not appear to affect this relationship or the average amount of oxygen consumed. No members of experimental group 2 were among the slowest runners.

---

## Pie Charts

To create a pie chart, select **Graphs** → **Pie Chart . . .**



**Figure 5.17.** Pie Chart Dialog

Select variables from the candidate list and click on the **Chart** button to produce a pie chart for each variable.

Select **2-D** or **3-D** under **Pie type** to specify whether you want to display a two-dimensional or three-dimensional chart.

---

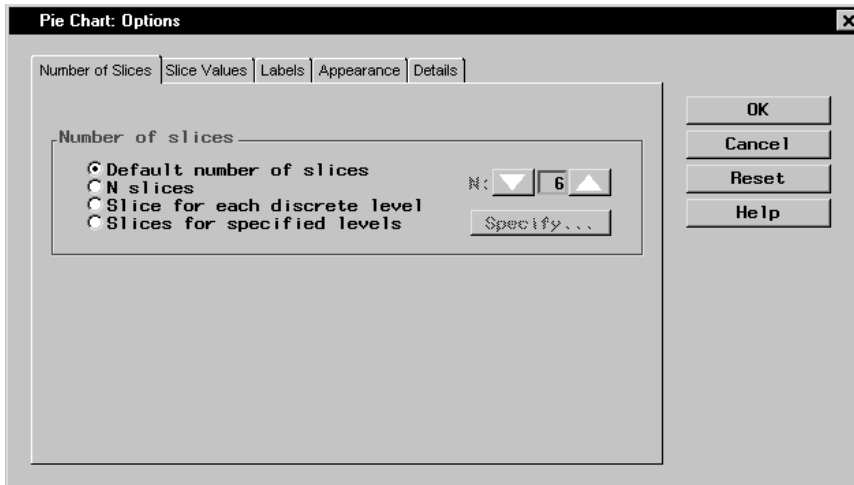
## Pie Chart Options

In the Pie Chart Options dialog, you can control the appearance of your pie chart. Click on the **Options** button to display the Pie Chart Options dialog. Click **OK** to save your changes.



## Number of Slices

The **Number of Slices** tab enables you to specify the number of slices in the chart and the levels for which they are displayed.

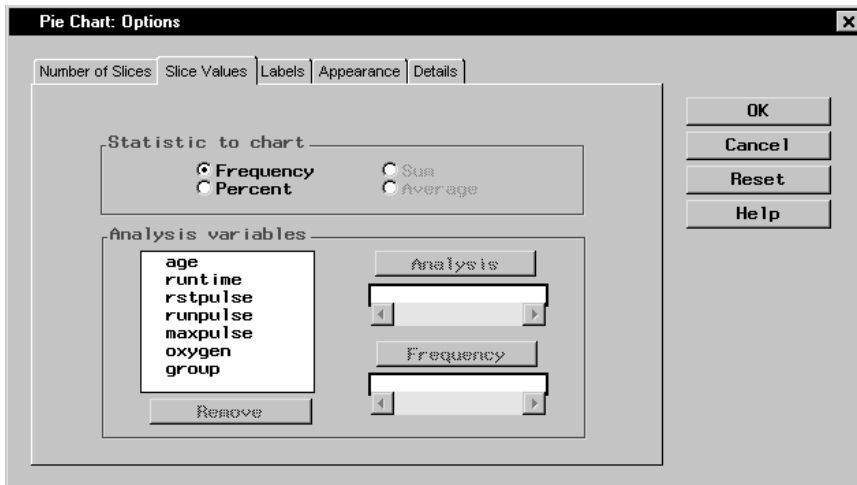


**Figure 5.18.** Number of Slice Tab

Under **Number of slices**, select **Default number of slices** to display an algorithmically determined number of slices. Select **N slices** and select a number from the list to specify the number of slices to be displayed. Select **Slice for each discrete level** to display a slice for each discrete level of data. If you are charting no more than one variable, select **Slices for specified levels** and click on the **Specify** button to provide a list of midpoints or to specify a range of numeric values, or to provide a list of character values.

## Slice Values

The **Slice Values** tab enables you to control the type of information that is displayed by each slice by specifying the statistic to display in the chart and any additional variable to use in computing the statistic.



**Figure 5.19.** Slice Values Tab

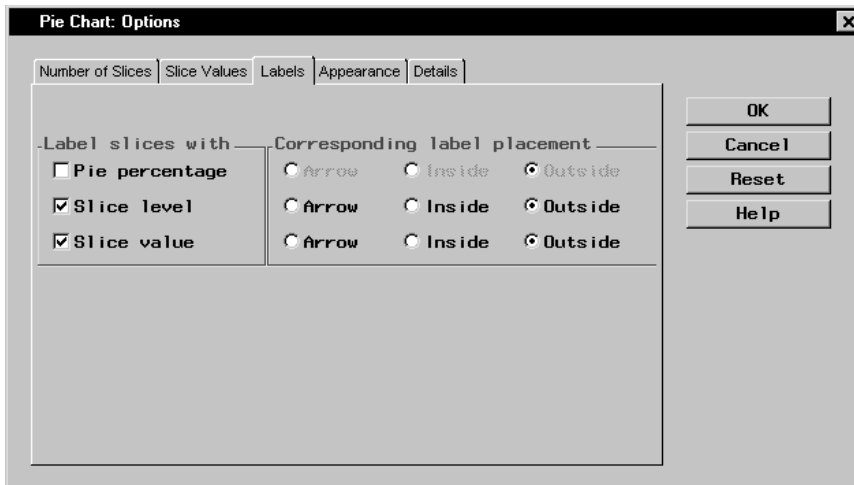
Selecting **Frequency** under **Statistic to chart** causes each slice to represent the frequency with which a value or range of values occurs for the chart variable. Selecting **Percent** causes each slice to represent the percentage of observations of the chart variable having a given value or falling into a given range.

If you want to show some characteristic of an additional variable for each level of the chart variable, select the additional variable as an **Analysis** variable. Then you can select **Sum** or **Average** of the analysis variable as the statistic to compute and display in each slice.

Select a **Frequency** variable if each observation in the data set represents several real observations, with values of the frequency variable indicating that number.

### Labels

The **Labels** tab enables you to define the labels for the slices in the pie chart.



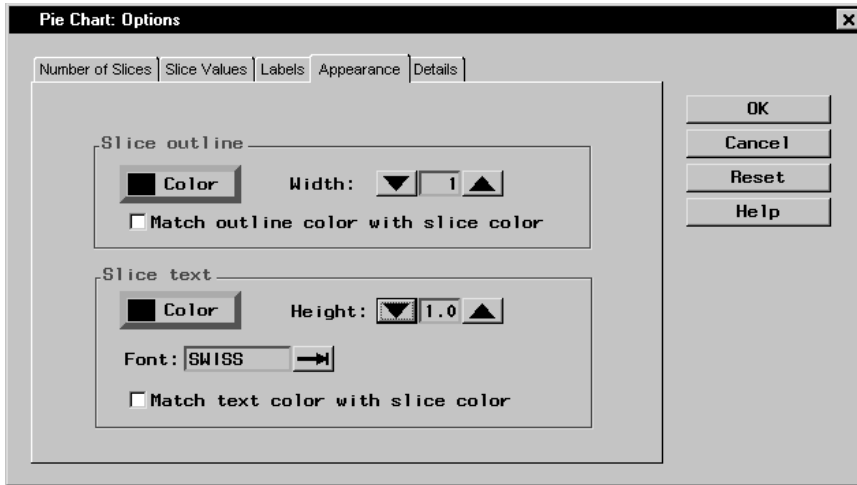
**Figure 5.20.** Labels Tab

Under **Label slices with**, you can choose to label the slices with their percentage of the pie chart, the level of the slice, and the value of the slice.

Under **Corresponding label placement**, you can place each of the labels inside or outside the slice, or you can include an arrow that points from the label to the slice.

### **Appearance**

The **Appearance** tab enables you to select colors, fonts, and line width.



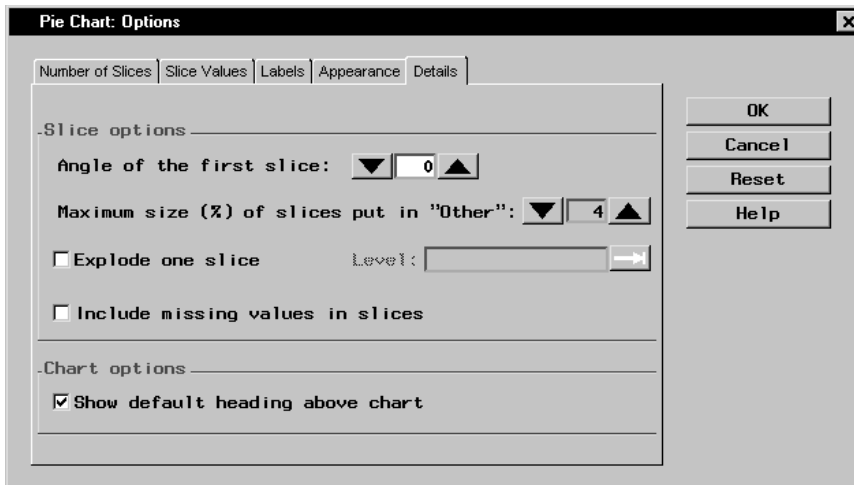
**Figure 5.21.** Appearance Tab

Under **Slice outline**, select the check box if you want the outline of each slice to be the same as the slice color. You can also control the width of the slice outlines. To select one color to be used for all outlines, click on the **Color** button and select a color from the Color Attributes dialog.

Under **Slice text**, select the check box if you want to match the color of the text with the color of the slice. You can also control the height and font of the slice text. To select one color to be used for all text, click on the **Color** button and select a color from the Color Attributes dialog.

### Details

The **Details** tab enables you to specify slice and chart heading options.



**Figure 5.22.** Details Tab

Under **Slice options**, you can specify the angle in degrees of the first slice by clicking on the up or down arrows or by typing in the degree. You can also define the maximum percentage size of slices you want to gather into an **Other** category by clicking on the arrows to choose from a range of one to fifteen percent. If you are charting one variable, you can select **Explode one slice**, and type in the level. If you have selected **Slice for each discrete level** or **Slices for specified levels** in the **Number of Slices** tab, you can click on the arrow next to **Level:** to select from a range of levels.

You can choose to include missing values in slices.

Under **Chart options**, you can select **Show default heading above chart** to include a heading that summarizes what the chart displays.

---

## Pie Chart Titles

Click on the **Titles** button to display the Titles dialog.

In the **Global** tab, you can specify titles that are displayed on all output. These titles are saved across Analyst sessions.

In the **Pie Chart** tab, you can specify titles for the pie chart. Select the box next to **Override global titles** to exclude the global titles from the pie chart results.

In the **Settings** tab, you can specify whether or not to include the date, page numbers, and a filter description.

---

## Pie Chart Variables

Click on the **Variables** button to display the Pie Chart Variables dialog.

BY group variables separate the data set into groups of observations. Separate analyses are performed for each group, and a separate chart is displayed for each analysis. For example, you could use a BY group variable to perform separate analyses on females and males. Specify BY group variables by selecting them in the candidate list and clicking on the **BY Group** button.

---

## Example: Create a 3-D Pie Chart

### *Open the Fitness Data Set*

In this example, you create a pie chart from the **Fitness** data set. If you have not already done so, open the **Fitness** data set by following these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select **Fitness**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Fitness** from the list of members.
7. Click **OK** to bring the **Fitness** data set into the data table.

### Specify Pie Chart Variable

To specify the variable to be charted and the chart type, follow these steps:

1. Select **Graphs** → **Pie Chart . . .**
2. Select **runtime** from the candidate list, and click **Chart** to make minutes to run 1.5 miles the charted variable.
3. Select **3-D** under **Pie type** to specify a three-dimensional chart.

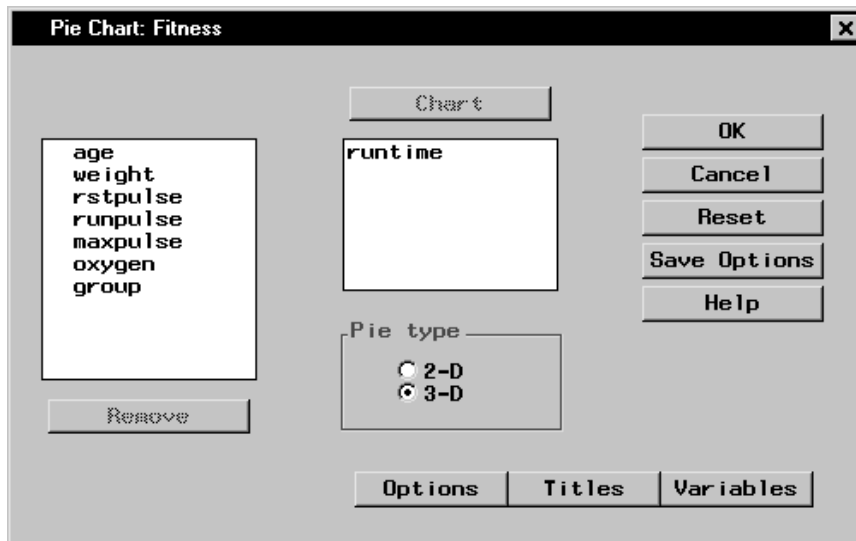
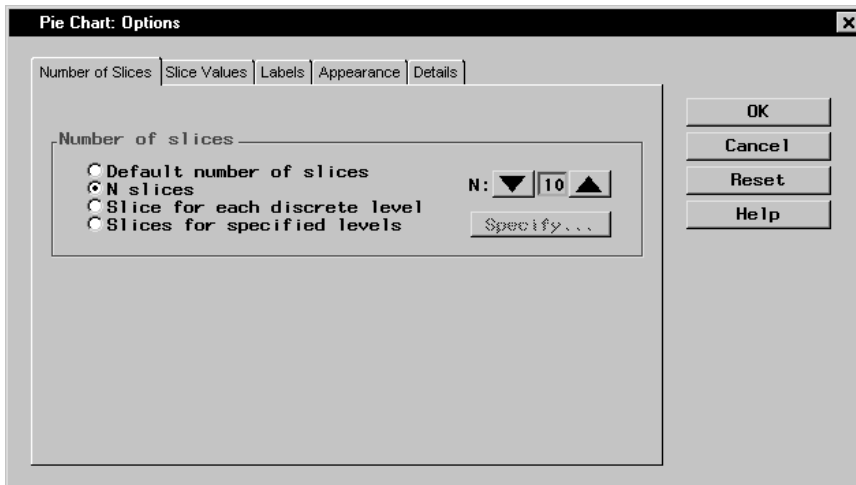


Figure 5.23. Pie Chart Variable and Type

### Specify Pie Chart Options

To specify your pie chart options, such as the number of slices, follow these steps:

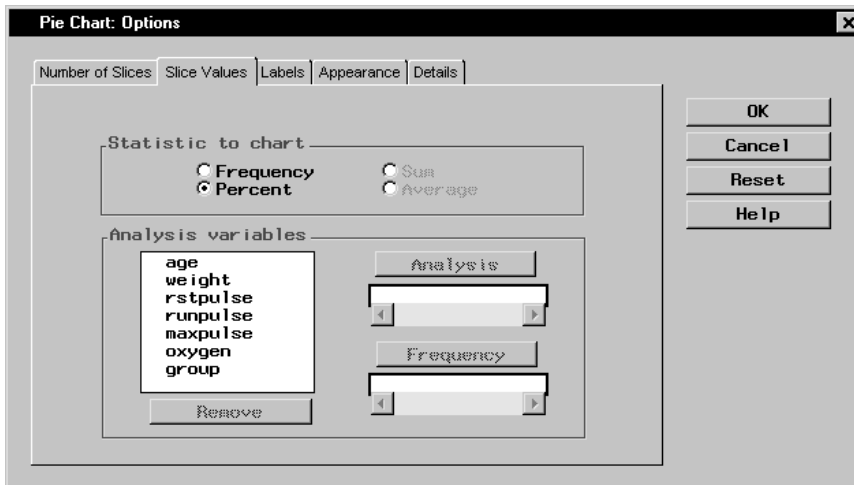
1. Click on the **Options** button to display the Pie Chart Options dialog.
2. In the **Number of Slices** tab, design a chart with ten slices by selecting **N slices** and clicking on the up arrow until the number **10** is visible.



**Figure 5.24.** Number of Slices in Pie Chart

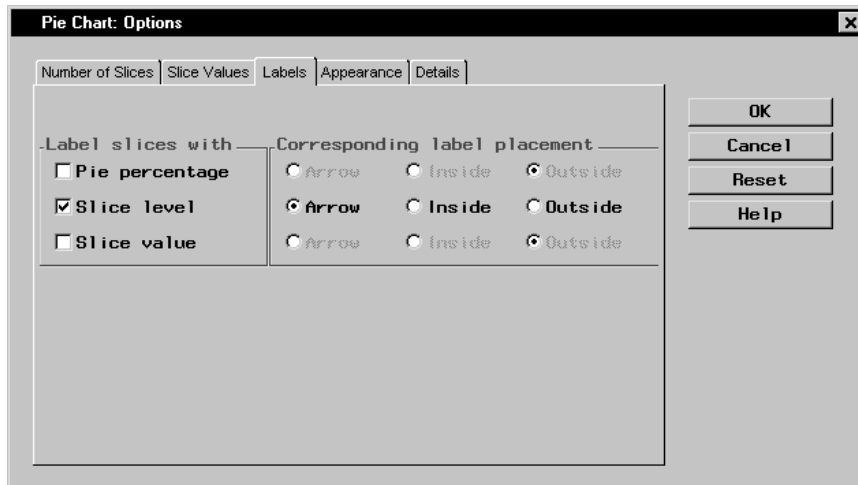
3. In the **Slice Values** tab, select **Percent** under **Statistic to chart** in order to chart the percentage of each runtime in relation to the whole.





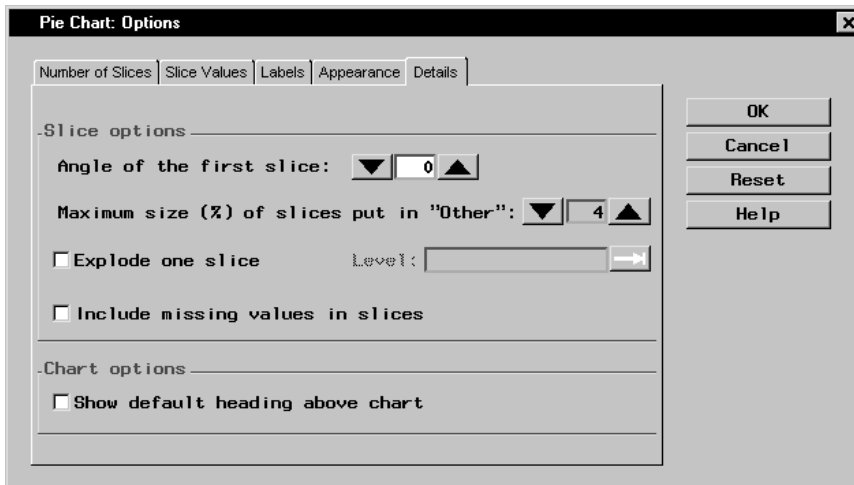
**Figure 5.25.** Statistic to Chart

4. In the **Labels** tab, select **Slice level** under **Label slices with**. Select **Arrow** under **Corresponding label placement**. Each slice indicates a runtime, and each label is placed outside the disc, with an arrow pointing to the corresponding slice.



**Figure 5.26.** Pie Chart Labels

5. In the **Details** tab, deselect **Show default heading above chart** under **Chart options**. You provide a new heading in the **Titles** dialog.



**Figure 5.27.** Deselect Default Heading

6. Click **OK** to save your changes and return to the Pie Chart dialog.

### **Specify Pie Chart Titles**

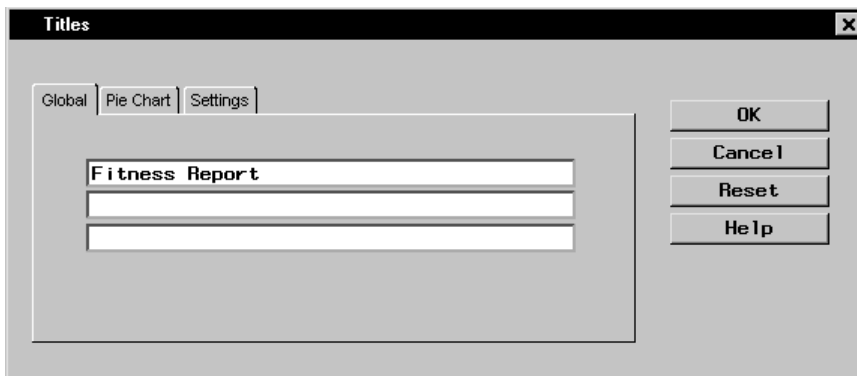
To specify the titles for your pie chart, follow these steps:

1. Click on the **Titles** button in the Pie Chart dialog.
2. In the **Pie Chart** tab, type **Percentage of Each Runtime** in the first field.



**Figure 5.28.** Pie Chart Title

3. If you did not change the global title in the first exercise in this chapter, click on the **Global** tab. Type **Fitness Report** in the first field. This global title is saved across all Analyst sessions until you change it.



**Figure 5.29.** Global Title

4. Click on **OK** to save your title changes.

### Generate Pie Chart

To display your pie chart, click **OK** in the Pie Chart dialog.

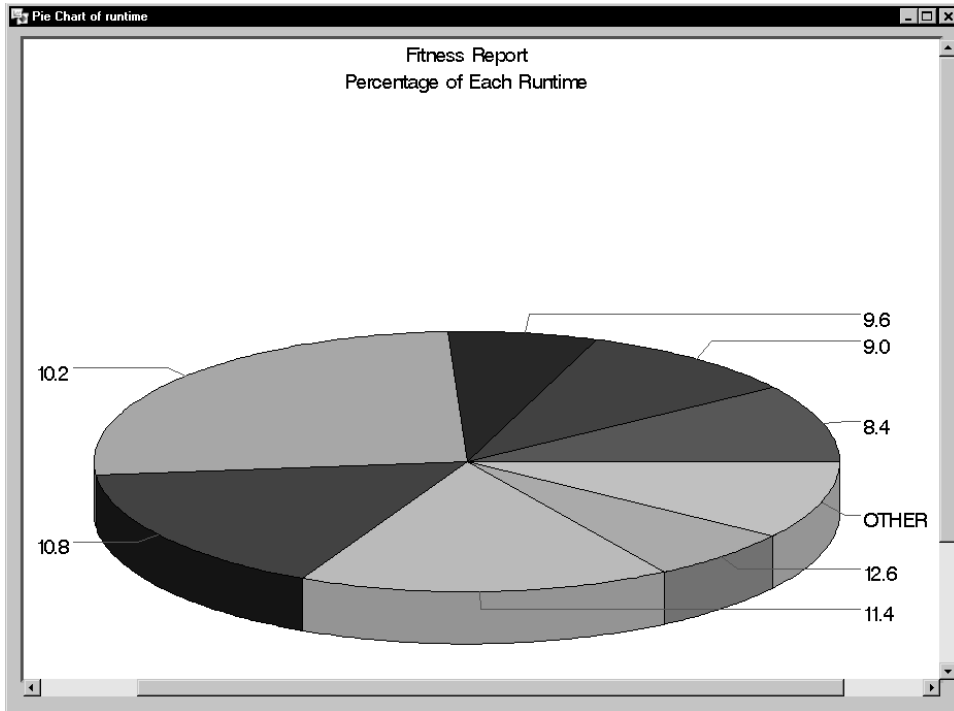
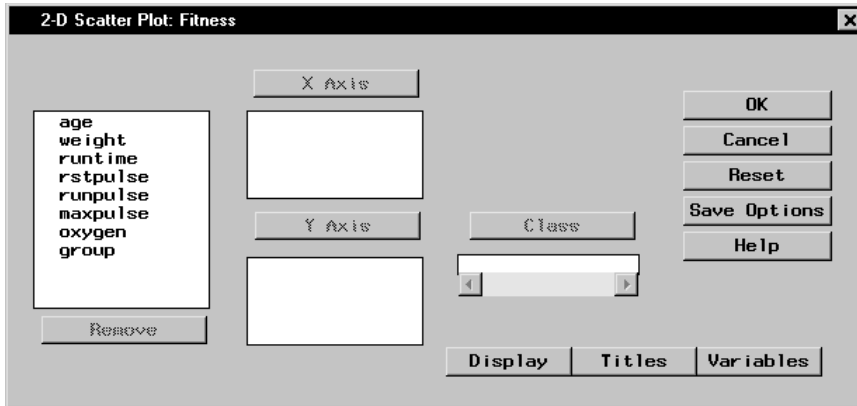


Figure 5.30. 3-D Pie Chart

---

## Scatter Plots

To create a scatter plot, select **Graphs** → **Scatter Plot**. Select **Two-Dimensional...** or **Three-Dimensional...** to create a two-dimensional or three-dimensional scatter plot of the data in the current table.



**Figure 5.31.** 2-D Scatter Plot Dialog

If you specify more than one variable for any of the axes, one plot is produced for each combination of variables.

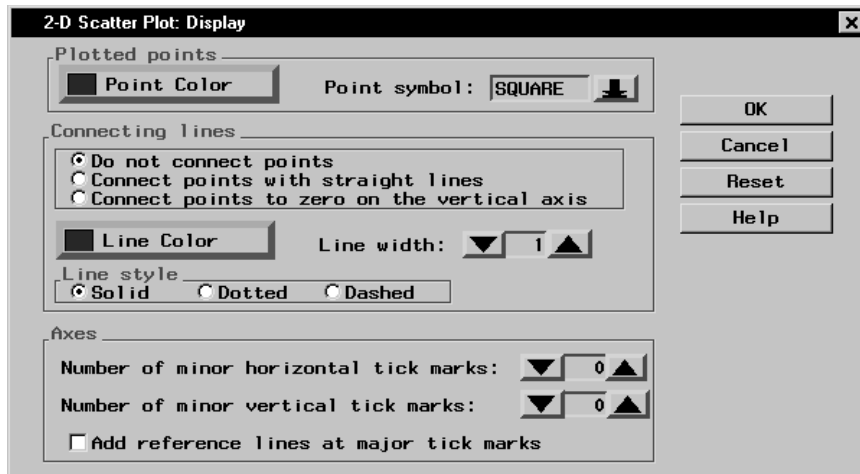
You must specify one or more  $x$ -axis variables and one or more  $y$ -axis variables. For three-dimensional plots, you must specify one or more  $z$ -axis variables.

For a two-dimensional scatter plot, specify a class variable to define subgroups. Each level of the class variable is represented by a different symbol on the scatter plot.

---

## Two-Dimensional Scatter Plot Options

In two-dimensional plots, you can specify the point color and connecting lines as well as control the tick marks on the axes. Click on the **Display** button to specify these display options.



**Figure 5.32.** 2-D Scatter Plot: Display Dialog

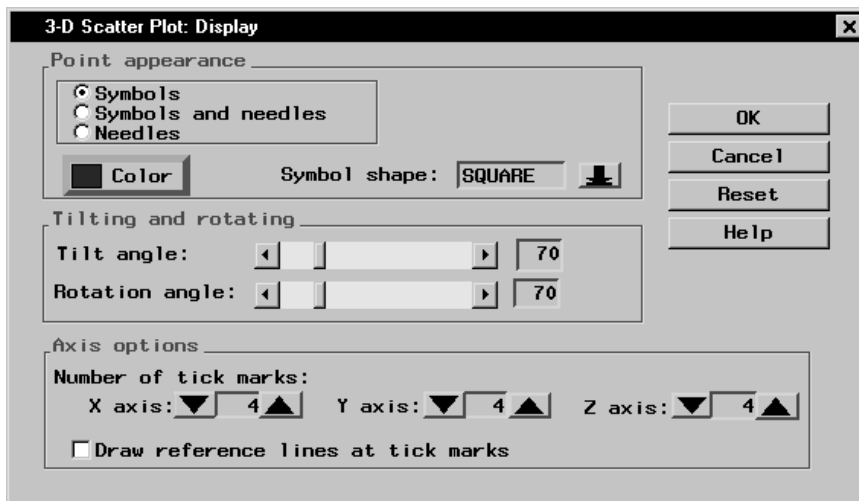
Click on the **Point Color** button to choose the point color. Click on the arrow next to **Point symbol:** to choose the symbol.

Under **Connecting lines**, specify whether the points are to be unconnected or connected to each other or the vertical axis, and specify the line color and style. Click on the **Line Color** button to specify the line color to be used for connecting points. Click on the arrows next to **Line width:** to specify the width of the line used to connect points. Under **Line style**, specify the style of the line used to connect points.

Under **Axes**, click on the up and down arrows to increase or decrease the number of minor horizontal and vertical tick marks. Select the check box to add reference lines at major tick marks.

## Three-Dimensional Scatter Plot Options

In three-dimensional plots, you can control the appearance of the points as well as the tilt and rotation of the plot. You can also control the tick marks on the axes.



**Figure 5.33.** 3-D Scatter Plot: Display Dialog

Under **Point appearance**, specify whether the points should be represented by symbols, needles, or both. Click on the **Color** button to specify the color for point symbols and needles. Click on the arrow next to **Symbol shape:** to specify the symbol for the points.

Under **Tilting and rotating**, move the bars next to **Tilt angle:** and **Rotation angle:** to specify the tilt angle and rotation angle for the plot.

Under **Axis options**, click on the arrows to specify the number of  $x$ -axis,  $y$ -axis, and  $z$ -axis tick marks. Click on the box next to **Draw reference lines at tick marks** to request that reference lines be drawn at each tick mark.



---

## Scatter Plot Titles

Click on the **Titles** button to display the Titles dialog.



**Figure 5.34.** Titles Dialog, 3-D Scatter Plot Tab

In the **Global** tab, you can specify titles that are displayed on all output. These titles are saved across Analyst sessions.

In the **Scatter Plot** tab, you can specify titles for the scatter plot. Select the box next to **Override global titles** to exclude the global titles from the scatter plot results.

In the **Settings** tab, you can specify whether or not to include the date, the page numbers, and a filter description.

---

## Scatter Plot Variables

Click on the **Variables** button to display the Scatter Plot Variables dialog.

BY group variables separate the data set into groups of observations. Separate analyses are performed for each group, and a separate plot is displayed for each analysis. For example, you could use a BY group variable to perform separate analyses on females and males. Specify BY group variables by selecting them in the candidate list and clicking on the **BY Group** button.

---

## Example: Create a 2-D Scatter Plot

### Open the Fitness Data Set

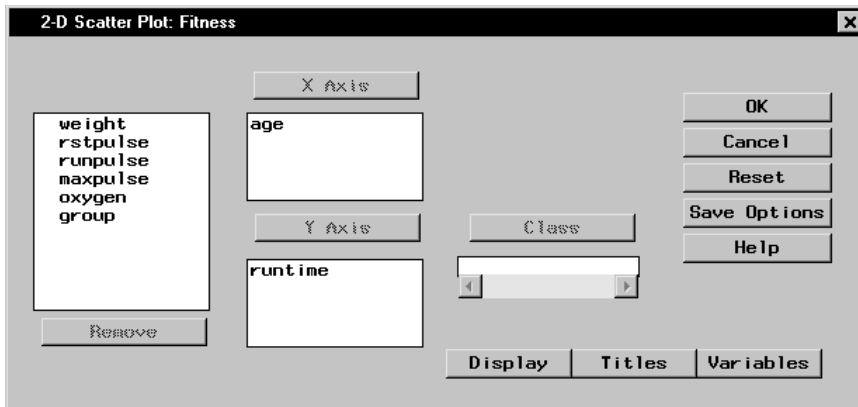
In this example, you use the `Fitness` data set as the basis of your scatter plot. If you have not already done so, open the `Fitness` data set by following these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select `Fitness`.
3. Click **OK** to create the sample data set in your `Sasuser` directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select `Sasuser` from the list of **Libraries**.
6. Select `Fitness` from the list of members.
7. Click **OK** to bring the `Fitness` data set into the data table.

### Specify Scatter Plot Variables

To specify the variables to be plotted, follow these steps:

1. Select **Graphs** → **Scatter Plot** → **Two-Dimensional** . . .
2. Select `age` from the candidate list, and click **X Axis** to make age in years the  $x$ -axis variable.
3. Select `runtime` from the candidate list, and click **Y Axis** to make minutes to run 1.5 miles the  $y$ -axis variable.

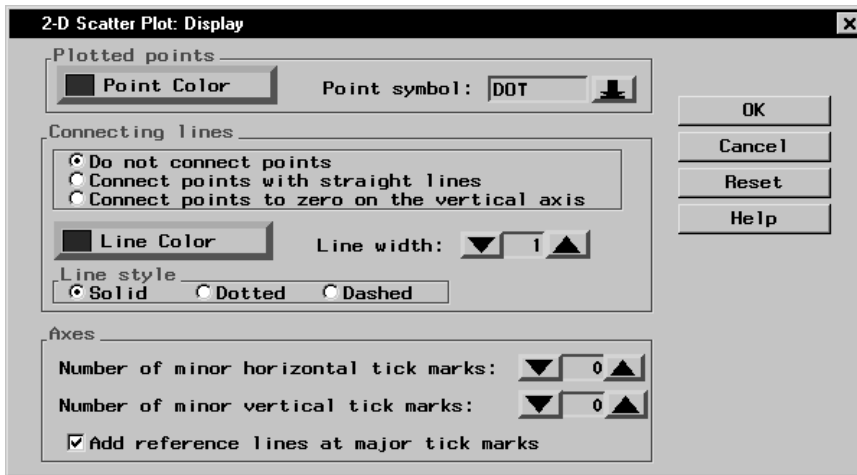


**Figure 5.35.** Scatter Plot Variables

### **Specify Scatter Plot Display Options**

To specify your scatter plot display options, follow these steps:

1. Click on the **Display** button to display the Scatter Plot Display dialog.
2. Under **Plotted points**, click on the **Point Color** button. Select **Red** from the list of colors to make your scatter plot points red. Click **OK**.
3. Click on the down arrow next to **Point symbol:** and select **DOT** from the list. This makes your scatter plot points display as dots.
4. Under **Axes**, select **Add reference lines at major tick marks**. This displays a grid on the scatter plot by which you can orient the points on the axes.



**Figure 5.36.** Display Options

5. Click **OK** to save your display changes.

### **Specify Scatter Plot Titles**

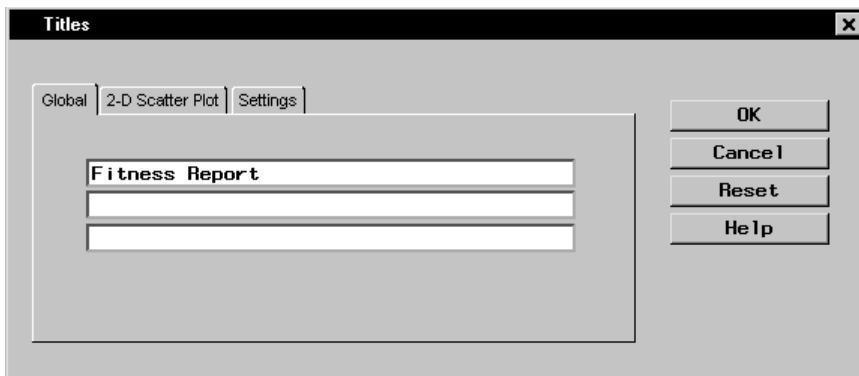
To specify the titles for your scatter plot, follow these steps:

1. Click on the **Titles** button in the Scatter Plot dialog.
2. In the **Scatter Plot** tab, type **Age versus Runtime** in the first field.



**Figure 5.37.** Scatter Plot Title

3. If you did not change the global title in the first exercise in this chapter, click on the **Global** tab. Type **Fitness Report** in the first field. This global title is saved across all Analyst sessions until you change it.

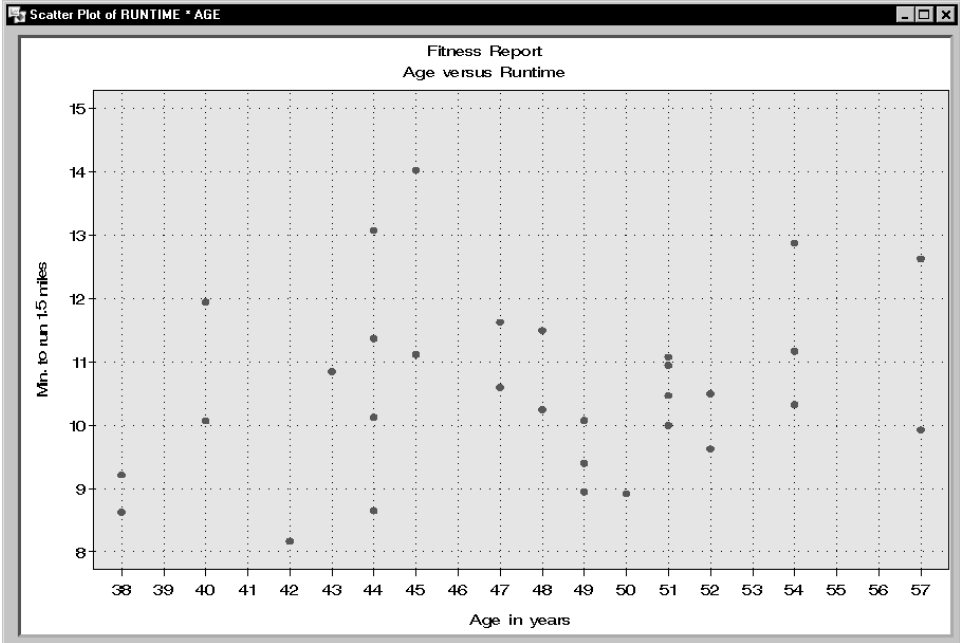


**Figure 5.38.** Global Title

4. Click **OK** to save your title changes.

**Generate Scatter Plot**

To display your scatter plot, click **OK** in the Scatter Plot dialog.



**Figure 5.39.** 2-D Scatter Plot

# Chapter 6

## Creating Reports

### Chapter Contents

---

<b>Introduction</b> . . . . .	153
<b>Listing Data</b> . . . . .	153
List Data Options . . . . .	154
List Data Titles . . . . .	155
List Data Variables . . . . .	156
Example: Create a Listing Report . . . . .	157
<b>Creating a Table</b> . . . . .	162
First Report Style . . . . .	162
Second Report Style . . . . .	165
Third Report Style . . . . .	166
Fourth Report Style . . . . .	167
Fifth Report Style . . . . .	168
Example: Create a Tabular Report . . . . .	169





# Chapter 6

## Creating Reports

---

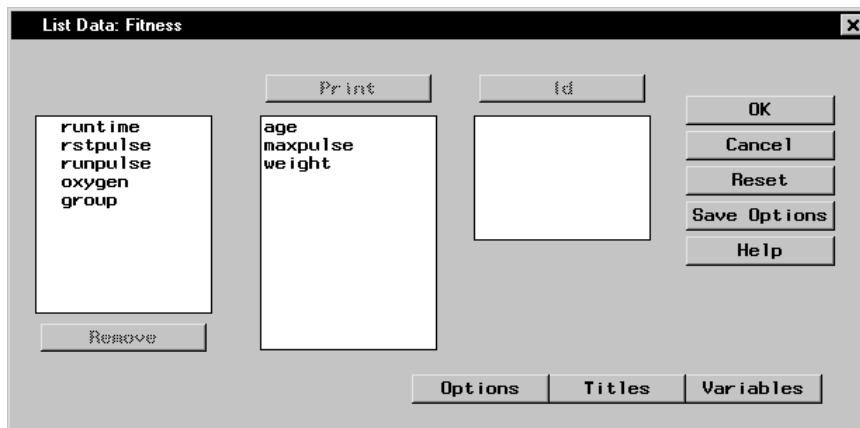
### Introduction

You can create a detailed report that lists portions of your data, or you can create a tabular report that summarizes your data.

---

### Listing Data

To create a detailed listing report, select **Reports** → **List Data . . .**



**Figure 6.1.** List Data Dialog

You can use the List Data dialog to print your data in a listing report. You can specify the variables to be included in the report and some details about the report format.

Select variables from the candidate list and click on the **Print** button to include the variables in the listing.

Select variables from the candidate list and click on the **Id** button to designate the variables as Id variables in the listing. These Id variables are used instead of observation numbers to identify the observations in the listing.

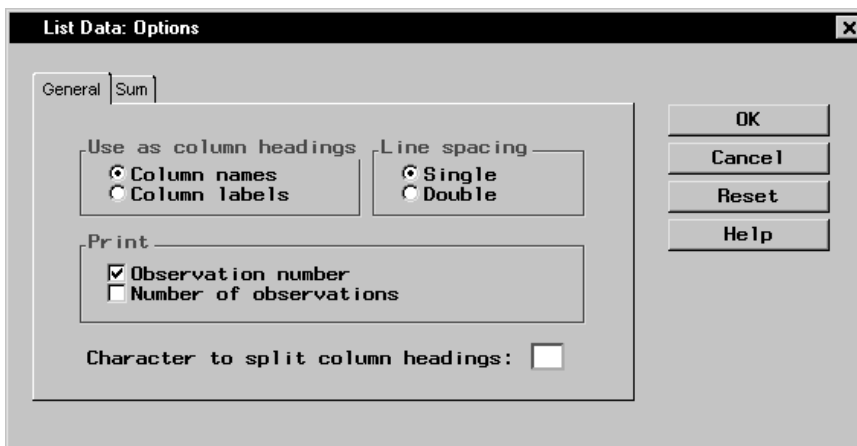
---

## List Data Options

Click on the **Options** button in the List Data dialog to specify options that control aspects of the report format and whether or not to print a sum for numeric columns.

### General

The **General** tab enables you to choose to use column names or column labels as column headings.



**Figure 6.2.** General Tab

Spacing between lines of the report can be single or double.

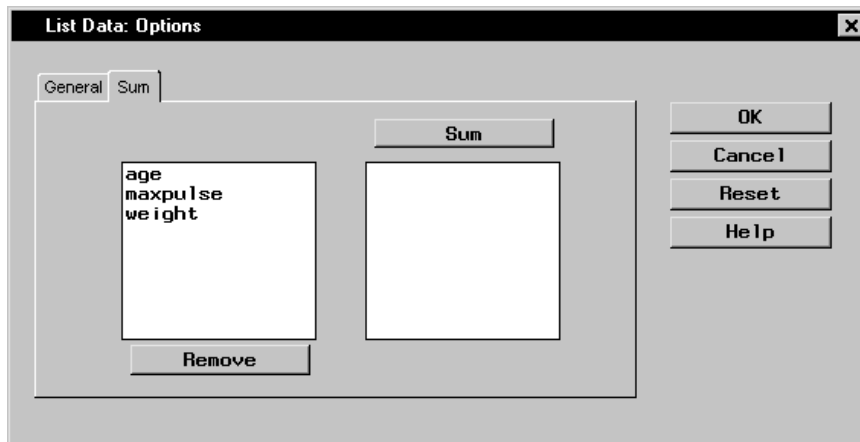
By default, you can print the number of each observation at the left as an identifier. If you have selected an Id variable, you cannot print the observation number.

You can also select to print the total number of observations in the data table at the end of the report.

To precisely control column headings in the report, you can specify a special character for variable labels that determines where the label is split as it forms a column heading. You can alter variable labels by selecting **Column Properties . . .** from the **Data** menu.

## Sum

The **Sum** tab enables you to generate a total for each selected numeric column.



**Figure 6.3.** Sum Tab

The numeric columns that are selected to be printed are listed in the candidate list. Select a column and click on the **Sum** button, or double-click on the column name to add it to the list of columns to be totalled.

---

## List Data Titles

Click on the **Titles** button to display the Titles dialog.



**Figure 6.4.** Titles Dialog, List Data Tab

In the **Global** tab, you can specify titles that are displayed on all output. These titles are saved across all Analyst sessions.

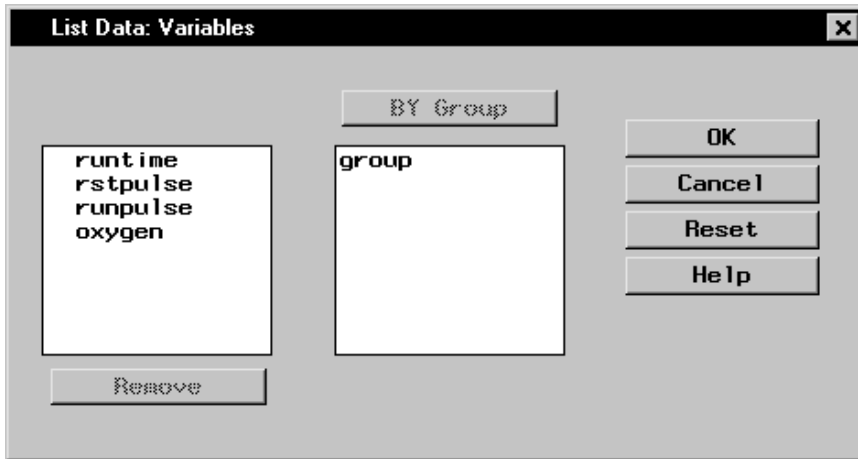
In the **List Data** tab, you can specify titles for the report. Select the box next to **Override global titles** to exclude the global titles from the report results.

In the **Settings** tab, you can specify whether or not to include the date, the page numbers, and a filter description.

---

## List Data Variables

Click on the **Variables** button to display the List Data: Variables dialog.



**Figure 6.5.** List Data: Variables Dialog

BY group variables separate the data set into groups of observations. Separate reports are produced for each group. For example, you could use a BY group variable to produce separate reports for females and males. Specify BY group variables by selecting them in the candidate list and clicking on the **BY Group** button.

---

## Example: Create a Listing Report

### *Open the Fitness Data Set*

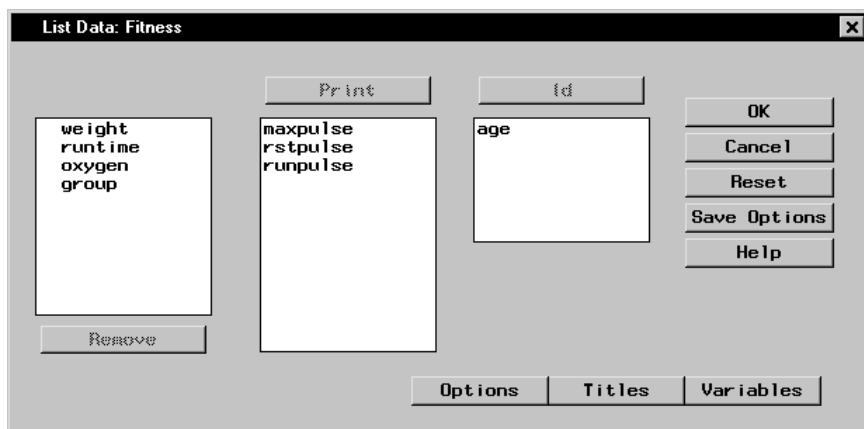
In this example, you use the Fitness data set as the basis of your listing report. To open the Fitness data set, follow these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select **Fitness**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Fitness** from the list of members.
7. Click **OK** to bring the **Fitness** data set into the data table.

### Specify Report Columns

To list maximum pulse, resting pulse, and average running pulse for each age, follow these steps:

1. Select **Reports** → **List Data . . .**
2. Select **maxpulse**, **rstpulse**, and **runpulse** and click on the **Print** button to include these variables in the report.
3. Select **age** and click on the **Id** button to make **age** the Id variable.

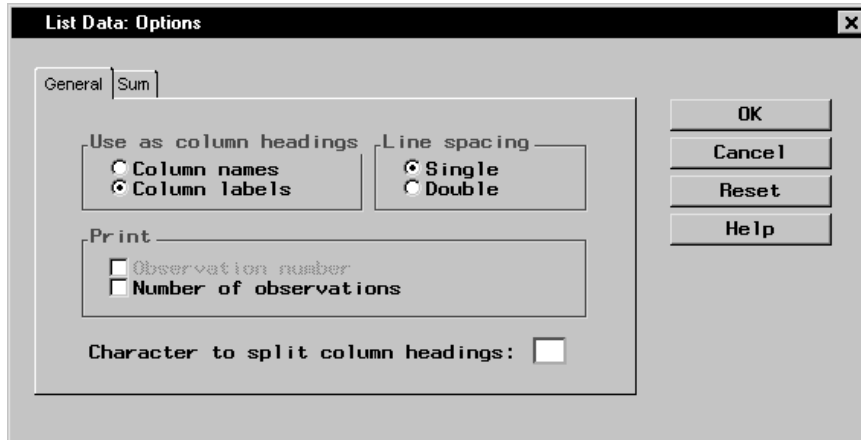


**Figure 6.6.** Columns in Report

## Specify Report Options

To designate options such as column headings, follow these steps:

1. Click on the **Options** button in the List Data dialog.
2. In the **General** tab, select **Column labels** under **Use as column headings**.



**Figure 6.7.** Use Column Labels as Column Headings

3. Click **OK** to save your changes.

## Specify Report Titles

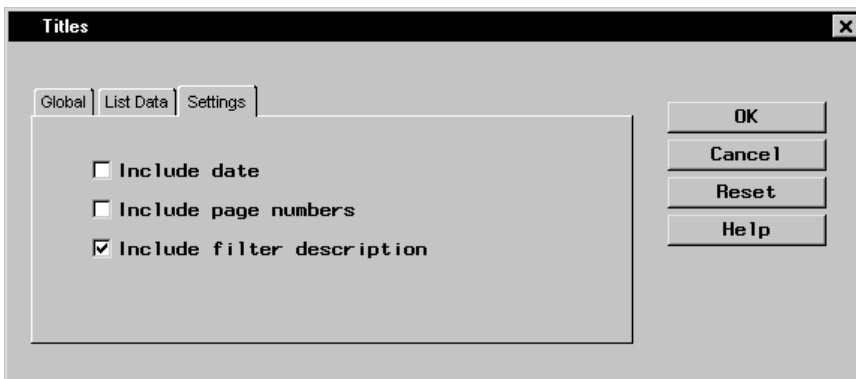
To specify the titles to be displayed in your report, follow these steps:

1. In the List Data dialog, click on the **Titles** button to specify your report titles.
2. In the **List Data** tab, type **Heart Rates According to Age** in the first field.



**Figure 6.8.** List Data Title

3. If you have not already done so, type **Fitness Report** in the first field in the **Global** tab.
4. Click on the **Settings** tab. Deselect **Include date** and **Include page numbers** so that the current date and page number are not printed on your report.



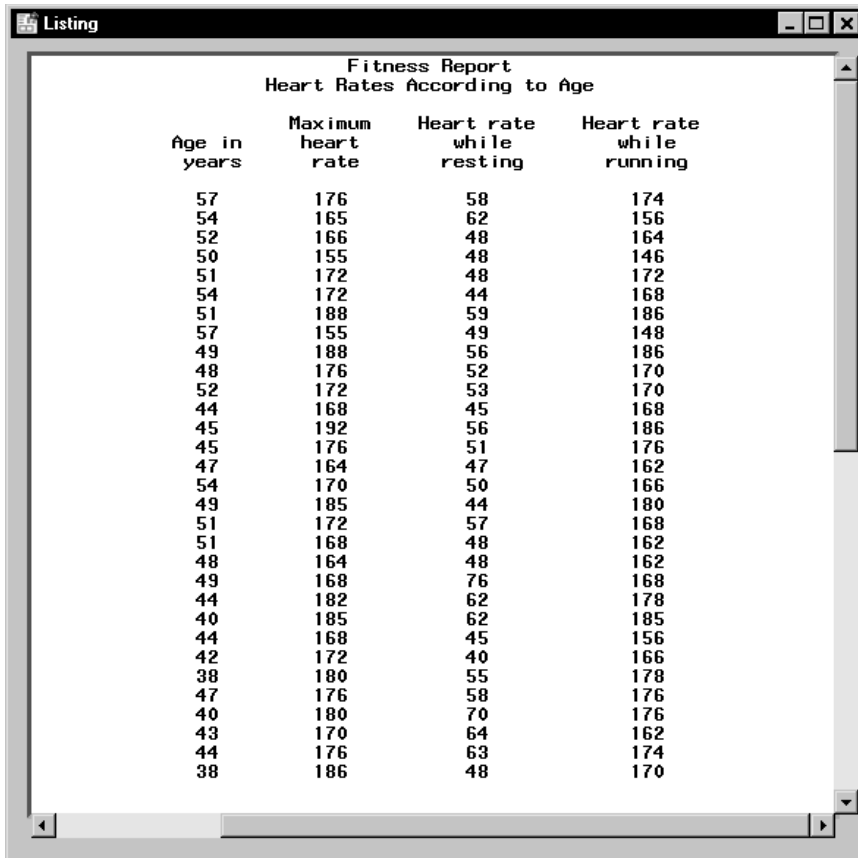
**Figure 6.9.** Exclude Date and Page Number

5. Click **OK** to save your title changes.



**Generate a Data Listing**

To generate a data listing of the columns that you have chosen, click **OK** in the List Data dialog.



The screenshot shows a window titled 'Listing' containing a table of fitness data. The table has four columns: 'Age in years', 'Maximum heart rate', 'Heart rate while resting', and 'Heart rate while running'. The data is presented in a list format with 30 rows of values.

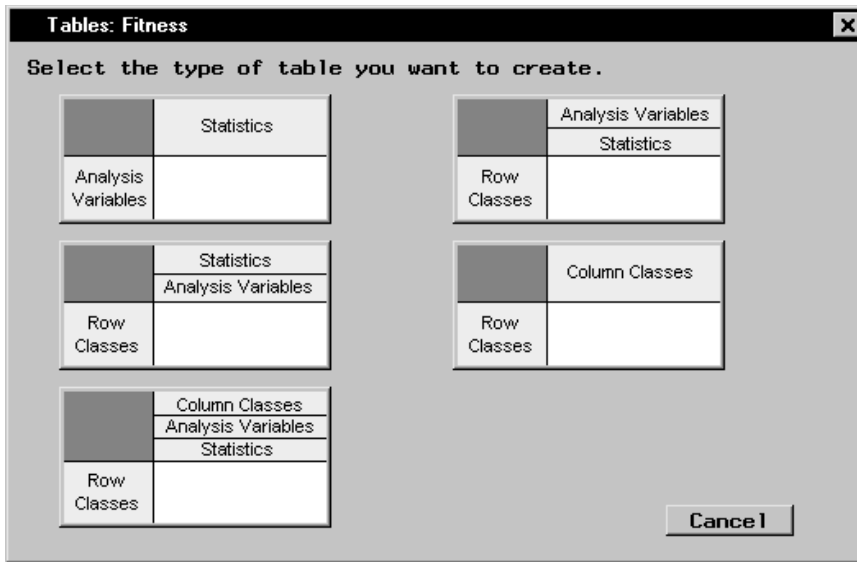
Age in years	Maximum heart rate	Heart rate while resting	Heart rate while running
57	176	58	174
54	165	62	156
52	166	48	164
50	155	48	146
51	172	48	172
54	172	44	168
51	188	59	186
57	155	49	148
49	188	56	186
48	176	52	170
52	172	53	170
44	168	45	168
45	192	56	186
45	176	51	176
47	164	47	162
54	170	50	166
49	185	44	180
51	172	57	168
51	168	48	162
48	164	48	162
49	168	76	168
44	182	62	178
40	185	62	185
44	168	45	156
42	172	40	166
38	180	55	178
47	176	58	176
40	180	70	176
43	170	64	162
44	176	63	174
38	186	48	170

**Figure 6.10.** Data Listing

## Creating a Table

A summary table can often help you spot important features of the data that are not apparent from a simple data listing.

To create a summary table, select **Reports** → **Tables . . .**



**Figure 6.11.** Reports Menu

Select a report style to specify the format and variables to be displayed.

### First Report Style

The first report style displays analysis variables as rows and statistics as columns.

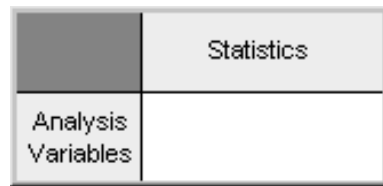


Figure 6.12. First Report Style

### Statistics

In the **Statistics** tab, select one or more statistics from the candidate list and click on the **Statistics** button to apply the statistics to the data in your report.

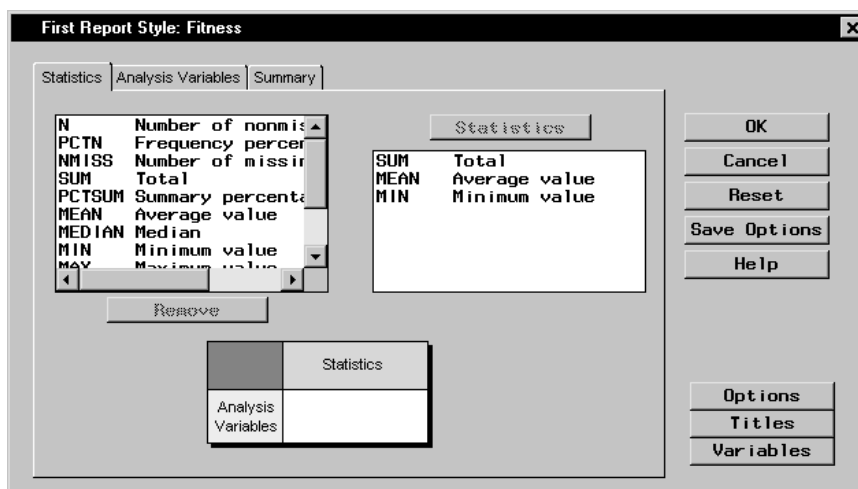
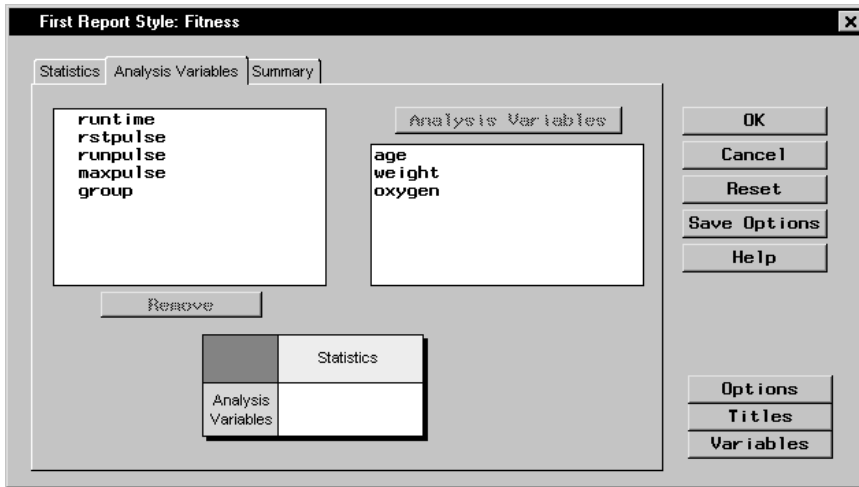


Figure 6.13. Statistics Tab

### Analysis Variables

An analysis variable is a variable for which statistics are computed. In the **Analysis Variables** tab, select one or more analysis variables from the candidate list and click on the **Analysis Variables** button to use these as analysis variables in your report.



**Figure 6.14.** Analysis Variables Tab

**Summary**

The **Summary** tab displays all of your selections. You can change the order of statistics and analysis variables by selecting the items in their lists and clicking the up and down arrows to change their position. Columns and rows in the resulting table are displayed in the tree view on the right.

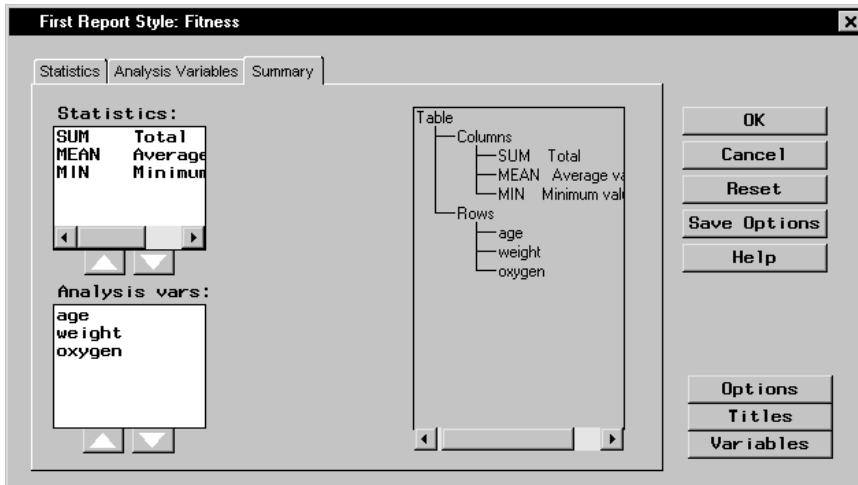


Figure 6.15. Summary Tab

## Second Report Style

The second report style displays levels of class variables as rows and statistics for analysis variables as columns.

	Analysis Variables
	Statistics
Row Classes	

Figure 6.16. Second Report Style

As with the first report style, the second report style also has **Statistics**, **Analysis Variables**, and **Summary** tabs. In addition, it also has a **Row Classes** tab.

### Row Classes

In the **Row Classes** tab, select one or more class variables from the candidate list and click on the **Row Classes** button to display rows in your report according to their levels.

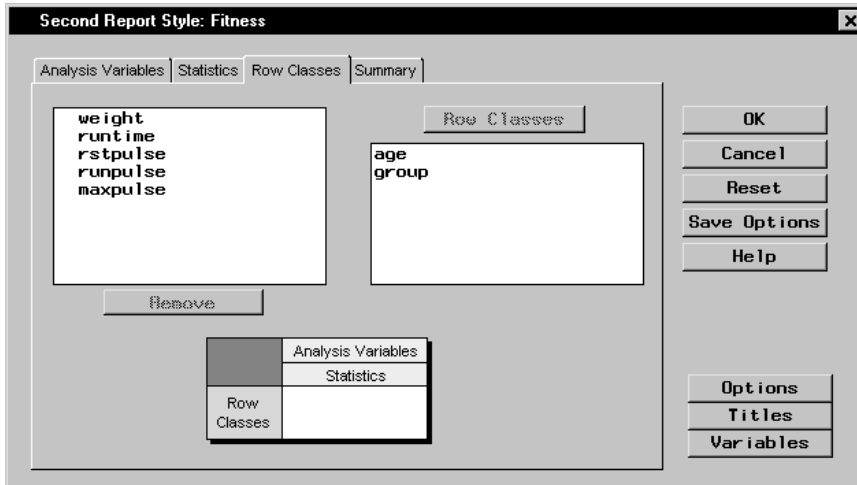


Figure 6.17. Row Classes Tab

### Third Report Style

The third report style displays levels of class variables as rows and statistics for analysis variables as columns.

	Statistics
	Analysis Variables
Row Classes	

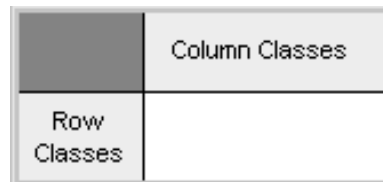
Figure 6.18. Third Report Style

The third report style contains the same tabs as the second report style; it differs from the second report style in the hierarchy of column headings.

---

## Fourth Report Style

The fourth report style displays levels of class variables as both rows and columns, with cells of the table containing the frequency of that combination of levels.



	Column Classes
Row Classes	

**Figure 6.19.** Fourth Report Style

As with the other report styles, the fourth report style has a **Summary** tab. As with the second and third report styles, the fourth report style has a **Row Classes** tab. In addition, this report style has a **Column Classes** tab.

### **Column Classes**

In the **Column Classes** tab, select one or more class variables from the candidate list and click on the **Column Classes** button to display columns in your report according to their levels.

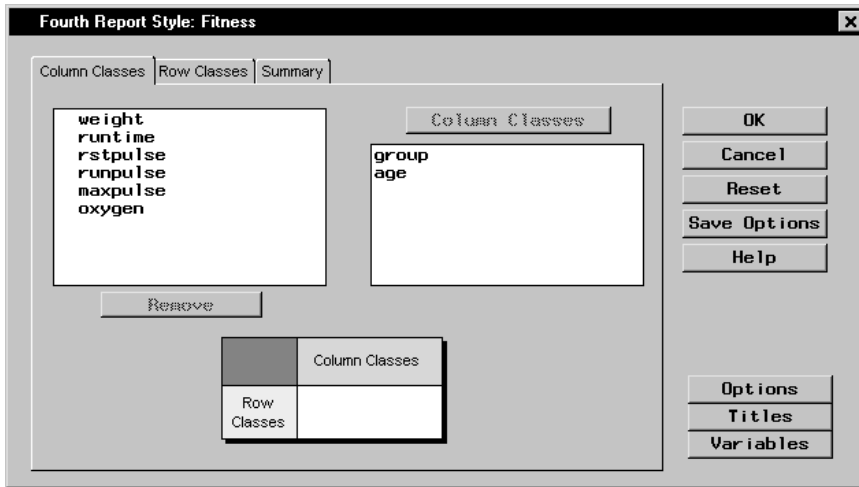


Figure 6.20. Column Classes Tab

## Fifth Report Style

The fifth report style displays levels of class variables as rows and statistics for analysis variables at levels of other class variables as columns.

	Column Classes
	Analysis Variables
	Statistics
Row Classes	

Figure 6.21. Fifth Report Style

As with other report styles, the fifth report style has a **Column Classes**, an **Analysis Variables**, a **Statistics**, a **Row Classes**, and a **Summary** tab.



---

## Example: Create a Tabular Report

### *Open the Class Data Set*

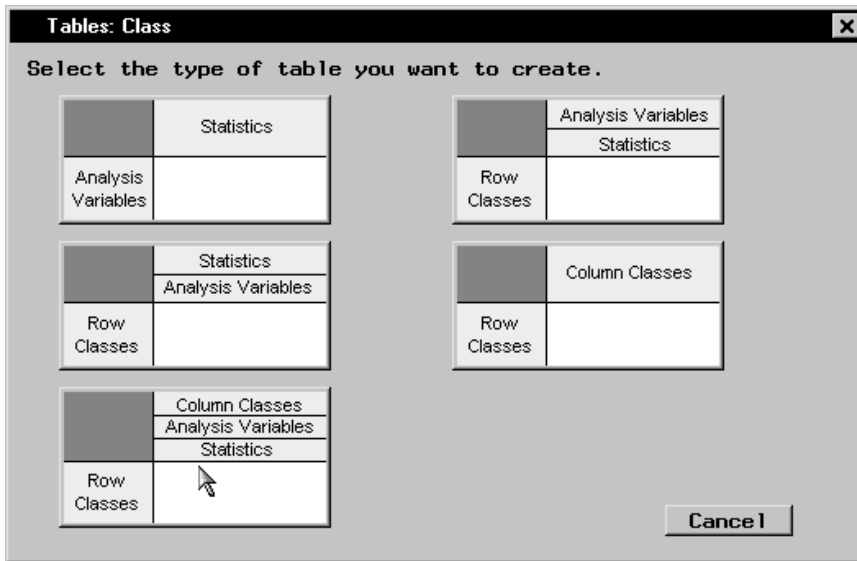
In this example, you use the **Class** data set as the basis of your report. To open the **Class** data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select **Class**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Class** from the list of members.
7. Click **OK** to bring the **Class** data set into the data table.

### *Choose a Report Style*

Use the fifth report style to display the average weights by age and sex in the **Class** data set. To choose a report style, follow these steps:

1. Select **Reports** → **Tables** . . .
2. Select the fifth report style.

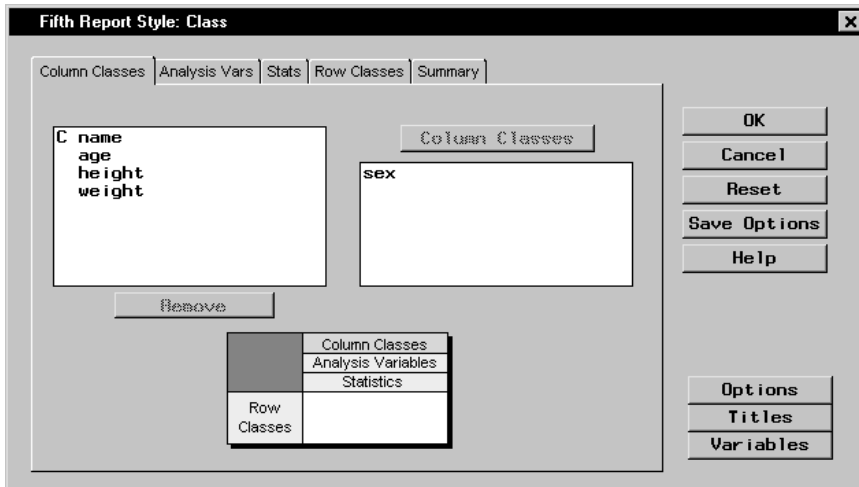


**Figure 6.22.** Select the Fifth Report Style

### **Specify Rows and Columns**

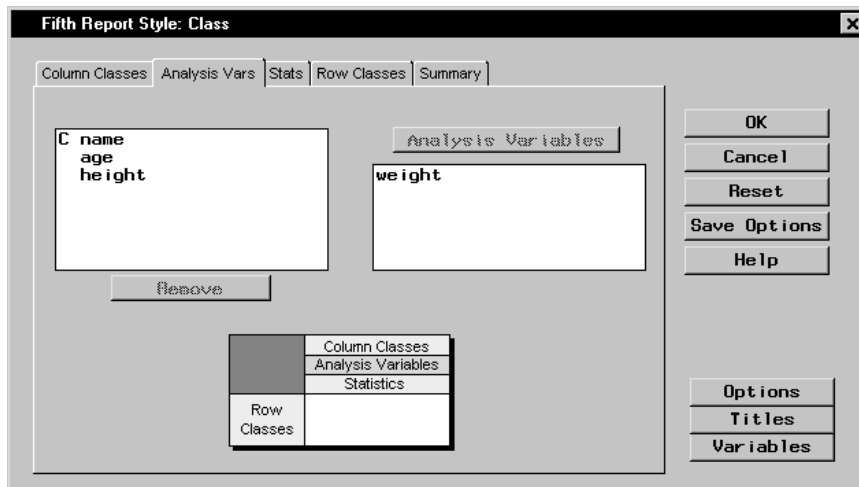
To specify the rows and columns for your report, follow these steps:

1. In the **Column Classes** tab, select **sex** from the candidate list and click on the **Column Classes** button to display the values of **sex** as columns in your report.



**Figure 6.23.** Select a Column Class

2. Click on the **Analysis Vars** tab. Select **weight** from the list and click on the **Analysis Variables** button to make **weight** the analysis variable in your report.



**Figure 6.24.** Select an Analysis Variable

3. Click on the **Stats** tab. Select **MEAN** from the list and click on the **Statistics** button to display the average weight.

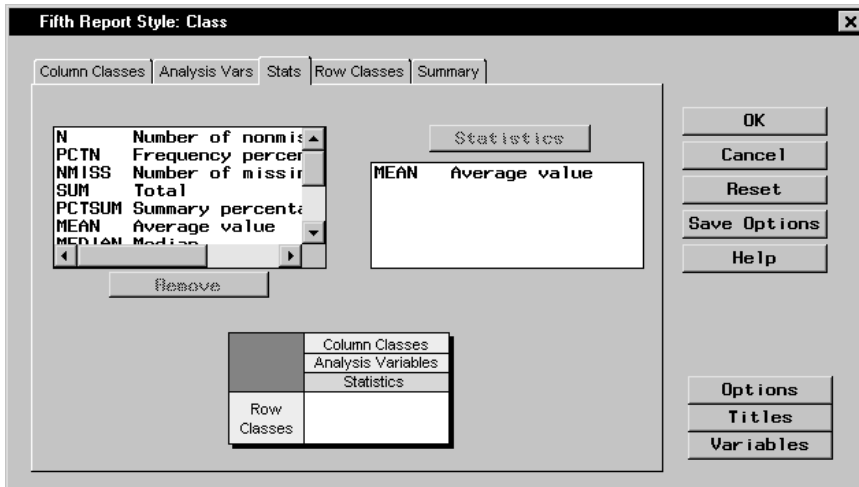
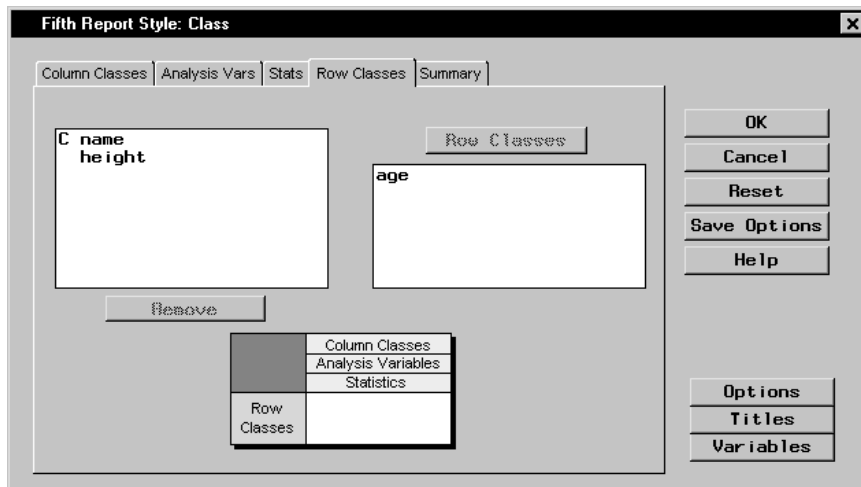


Figure 6.25. Select a Statistic

4. Click on the **Row Classes** tab. Select **age** from the list and click on the **Row Classes** button to display the values of **age** as the rows in your report.



**Figure 6.26.** Select a Row Class

5. Click on the **Summary** tab to see the results of your selections.

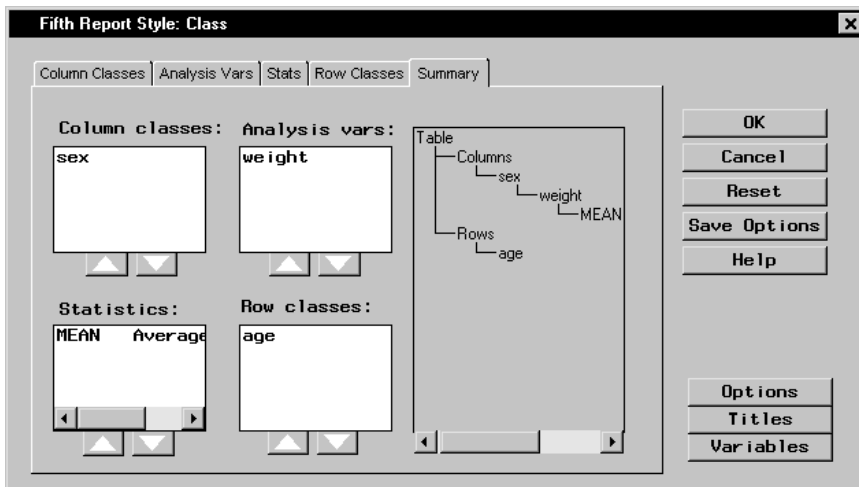
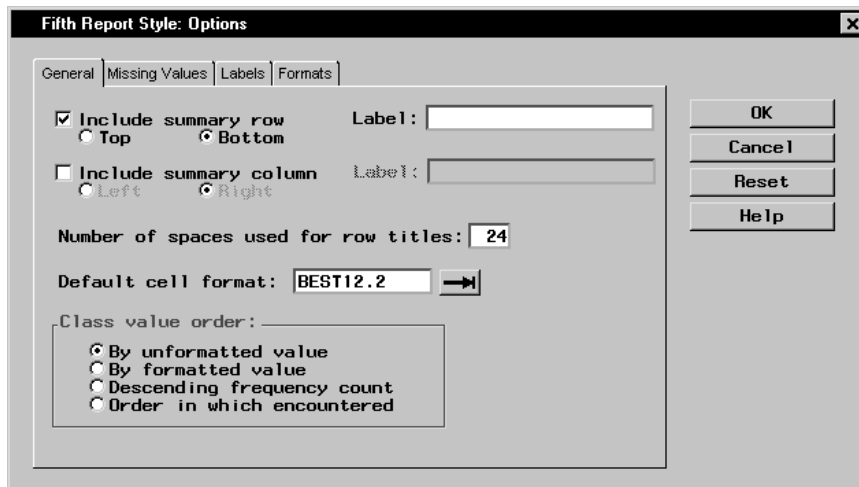


Figure 6.27. Report Layout

### Specify Report Options

To specify the options for your report, follow these steps:

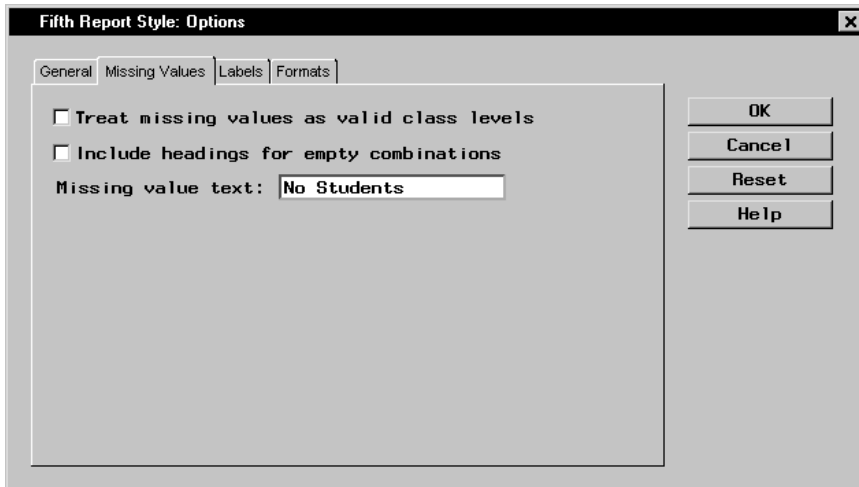
1. Click on the **Options** button in the Fifth Report Style dialog.
2. In the **General** tab, select **Include summary row**. Click **Bottom** to display a summary row at the bottom of each column.



**Figure 6.28.** Include Summary Row

3. Click on the **Missing Values** tab. Type **No Students** in the **Missing value text:** field.





**Figure 6.29.** Type Missing Value Text

4. Click **OK** to save your changes and return to the Fifth Report Style dialog.

### **Specify Report Titles**

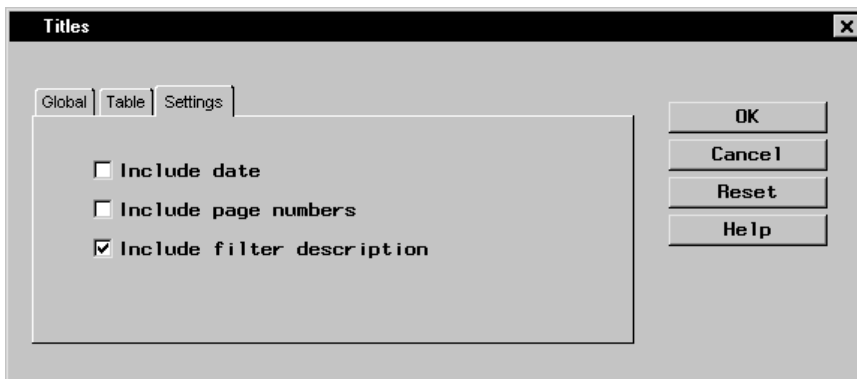
To create a title and suppress the date and page numbers in your report, follow these steps:

1. Select the **Titles** button in the Fifth Report Style dialog.
2. In the **Table** tab, type **Average Weights by Age and Sex** in the first field.
3. Select **Override global titles** to suppress the title from the previous example.



**Figure 6.30.** Add a Title

4. Click on the **Settings** tab. Deselect **Include date** and **Include page numbers** so that the date and page numbers are not displayed in your report.

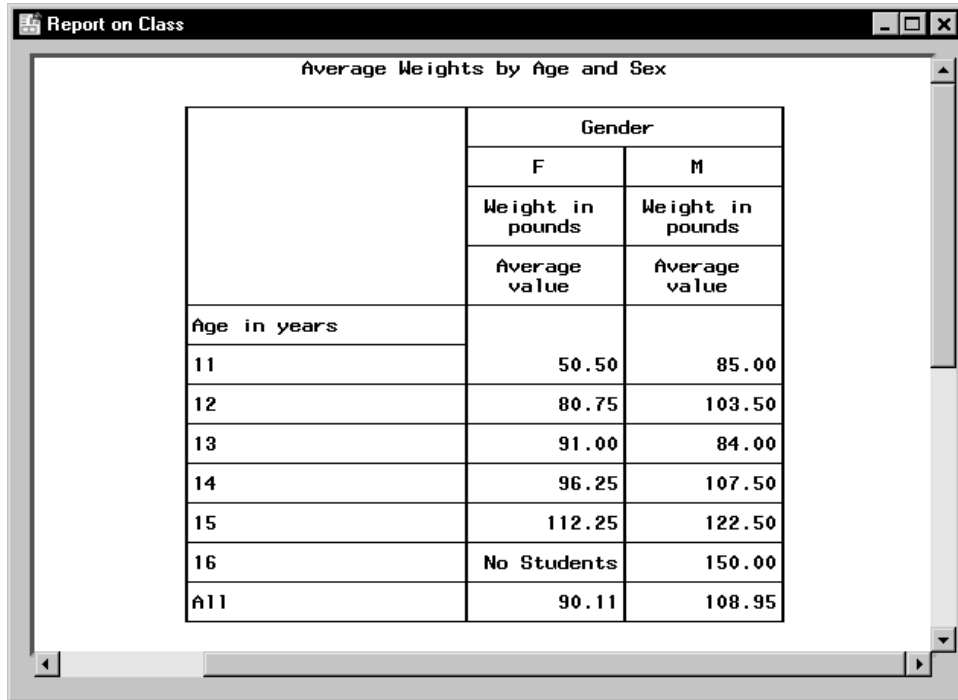


**Figure 6.31.** Suppress Date and Page Numbers

5. Click **OK** to save your changes and to return to the Fifth Report Style dialog.

### Display Your Report

To display your report in the fifth report style, click **OK** in the Fifth Report Style dialog.



The screenshot shows a window titled "Report on Class" with a scrollable area containing a table. The table is titled "Average Weights by Age and Sex". The table has three columns: "Age in years", "F", and "M". The "F" and "M" columns are further divided into "Weight in pounds" and "Average value". The data rows show average weights for ages 11 through 16, and an "All" row. The "F" column for age 16 shows "No Students".

Age in years	Gender	
	F	M
	Weight in pounds	Weight in pounds
	Average value	Average value
11	50.50	85.00
12	80.75	103.50
13	91.00	84.00
14	96.25	107.50
15	112.25	122.50
16	No Students	150.00
All	90.11	108.95

**Figure 6.32.** Display Report in Fifth Report Style



# Chapter 7

## Descriptive Statistics

### Chapter Contents

---

<b>Introduction</b> . . . . .	183
<b>Producing One-Way Frequencies</b> . . . . .	184
<b>Computing Summary Statistics</b> . . . . .	190
<b>Examining the Distribution</b> . . . . .	195
<b>Computing Correlations</b> . . . . .	200
<b>References</b> . . . . .	208



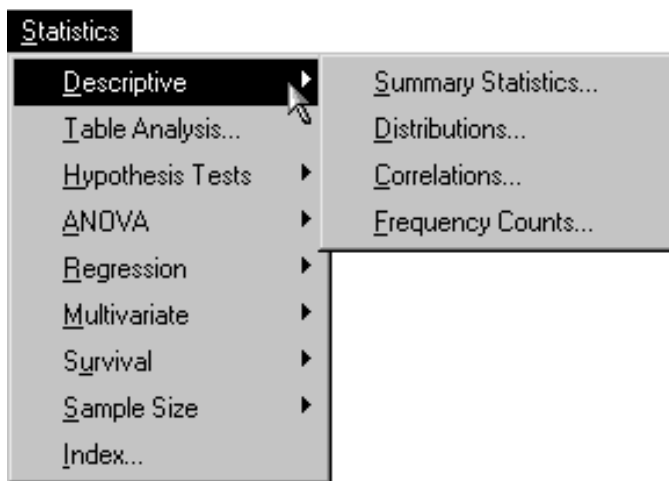
# Chapter 7

## Descriptive Statistics

---

### Introduction

Descriptive statistics and plots are often used in the initial phase of a statistical analysis. These tools enable you to identify relationships in the data and to determine directions for further analysis.



**Figure 7.1.** Descriptive Menu

The Analyst Application provides several types of descriptive statistics and graphical displays. The Summary Statistics task provides the following information: mean, median, standard error and standard deviation, variance, minimum, maximum, range, sum, skewness and kurtosis, student's  $t$  and probability value, coefficient of variation, and sums of squares. Graphics in this task include histograms and box-and-whisker plots.

The Distributions task produces statistics such as moments and quantiles as well as measures of location and variability. You can request fitted distributions from the normal, lognormal, Weibull, and exponential distributions. Plots included are the box-and-whisker plot, histogram, probability plot, and quantile-quantile plots. Histograms can be superimposed with fitted curves from the distribution families. Probability and quantile-quantile plots are available for each of the distributions.

The Correlations task gives you the choice of Pearson and Spearman correlations as well as Cronbach's alpha, Kendall's tau-*b*, and Hoeffding's *D*. Scatter plots with optional confidence ellipses are available.

The Frequency Counts task provides one-way frequency tables, which include frequencies, percentages, and cumulative frequencies and percentages. Horizontal and vertical bar charts are also available.

The examples in this chapter demonstrate how you can use the Analyst Application to compute one-way frequency tables, obtain summary statistics, examine the distribution of your data, and compute correlations.

---

## Producing One-Way Frequencies

The data set analyzed in the following sections is taken from the 1995 Statistical Abstract of the United States. The data are measures of the birth rate and infant mortality rate for 1992 in the United States. Information is provided for the 50 states and the District of Columbia. The states are grouped by region. Here, these data are considered to be a sample of yearly data.

Suppose you want to determine the frequency of occurrence of the various regions. In the following example, a listing of the frequencies and a bar chart are produced.

In the Frequency Counts task, you can compute one-way frequency tables for the variables in your data set. For each value of your analysis variable, Analyst produces the frequency, cumulative frequency, and cumulative percentage. You can control the order in which the values appear and specify group and count variables.



### **Open the Bthdth92 Data Set**

The data are provided in the Analyst Sample Library. To open the Bthdth92 data set, follow these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select Bthdth92.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select Sasuser from the list of **Libraries**.
6. Select Bthdth92 from the list of members.
7. Click **OK** to bring the Bthdth92 data set into the data table.

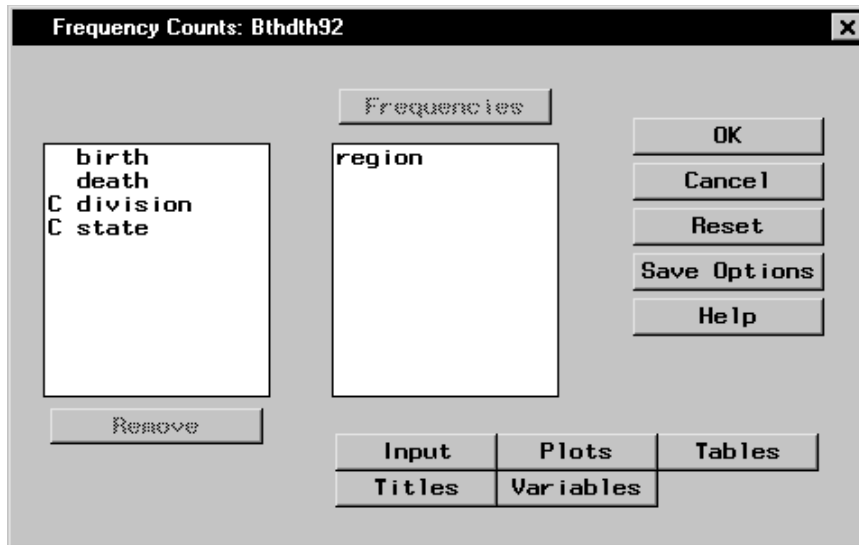
### **Request Frequency Counts**

To request frequency counts, follow these steps:

1. Select **Statistics** → **Descriptive** → **Frequency Counts . . .**
2. Select region as the frequencies variable from the candidate list.

The default analysis provides the information desired. Note that you can use the Input dialog to select the specific ordering by which the variable values are listed.

Figure 7.2 displays the Frequency Counts dialog with region specified as the frequencies variable.

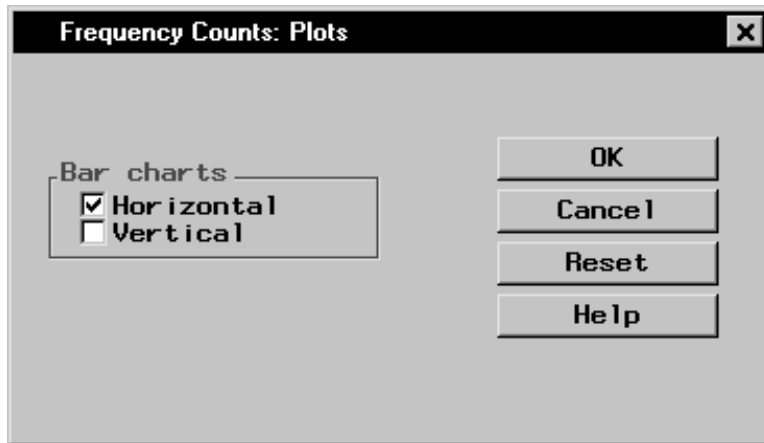


**Figure 7.2.** Frequency Counts Dialog

### ***Request a Horizontal Bar Chart***

To produce a horizontal bar chart in addition to the frequency counts, follow these steps:

1. Click on the **Plots** button.
2. Select **Horizontal**, as displayed in [Figure 7.3](#).
3. Click **OK** to close the Plots dialog.

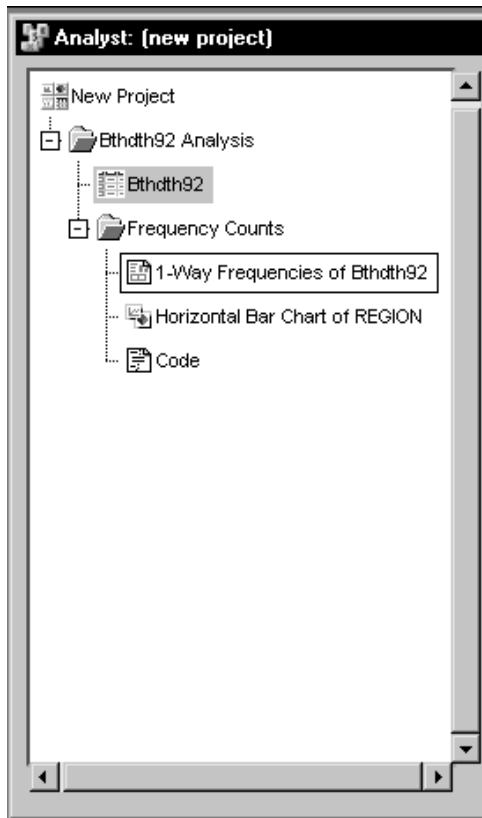


**Figure 7.3.** Frequency Counts: Plots Dialog

Click **OK** in the Frequency Counts main dialog to perform the analysis.

### ***Review the Results***

The results are presented in the project tree under the **Frequency Counts** folder, as displayed in [Figure 7.4](#). The three nodes represent the frequency counts output, the horizontal bar chart, and the SAS programming statements (labeled **Code**) that generate the output.



**Figure 7.4.** Frequency Counts: Project Tree

You can double-click on any node in the project tree to view the contents in a separate window. Note that the first output generated is displayed by default.

Figure 7.5 displays the table of frequency counts for the variable `region`.

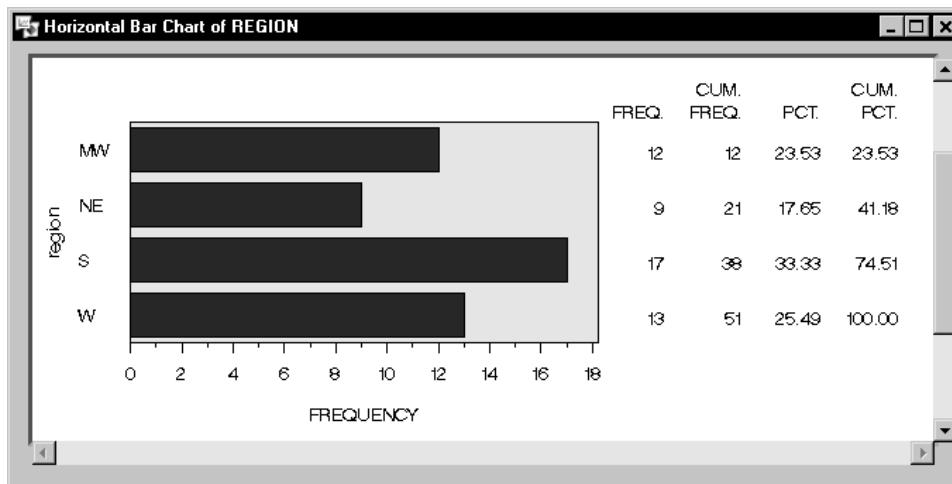
The screenshot shows a window titled "1-Way Frequencies of Bthdth92". Inside the window, the text "The FREQ Procedure" is centered above a table. The table has five columns: "region", "Frequency", "Percent", "Cumulative Frequency", and "Cumulative Percent". The rows represent the regions: MW, NE, S, and W. The cumulative values are calculated as the sum of the previous rows' values.

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
MW	12	23.53	12	23.53
NE	9	17.65	21	41.18
S	17	33.33	38	74.51
W	13	25.49	51	100.00

**Figure 7.5.** Frequency Counts: One-Way Frequencies of the Variable region

The table shows that about 33% of the observations in the data set are located in the southern region, and roughly 25% of the observations are located in the western and midwestern regions, respectively. Approximately 18% of the observations are located in the northeastern region.

To display the bar chart of the frequency counts, double-click the node labeled **Horizontal Bar Chart of REGION** (Figure 7.6).



**Figure 7.6.** Frequency Counts: Horizontal Bar Chart by Region

## Computing Summary Statistics

In this task, summary statistics (such as the mean, standard deviation, and minimum and maximum values) are desired for the birth and infant mortality rates for each region. In addition, box-and-whisker plots are requested.

### Request Summary Statistics

To request the Summary Statistics task, follow these steps:

1. Select **Statistics** → **Descriptive** → **Summary Statistics...**
2. Select the analysis variables **birth** and **death** from the candidate list.

You can specify a classification variable to define groups within your data. When you specify a classification variable, the Analyst Application produces summary statistics for the analysis variables at each level of the classification variable.

3. Select **region** as the classification variable.

Figure 7.7 displays the Summary Statistics main dialog with birth and death specified as the analysis variables and region specified as the classification variable.

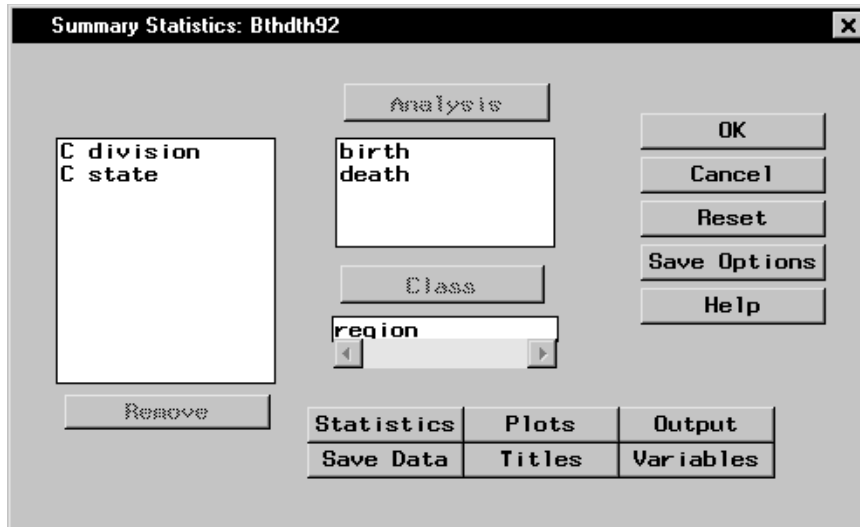


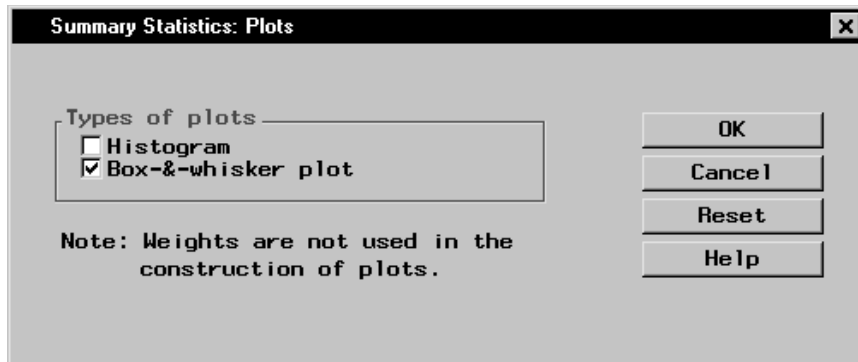
Figure 7.7. Summary Statistics Dialog

### **Request Box-and-Whisker Plots**

To request box-and-whisker plots, follow these steps:

1. Click on the **Plots** button.
2. Select **Box-&-whisker plot**.
3. Click **OK**.

Figure 7.8 displays the Plots dialog with **Box-&-whisker plot** selected.



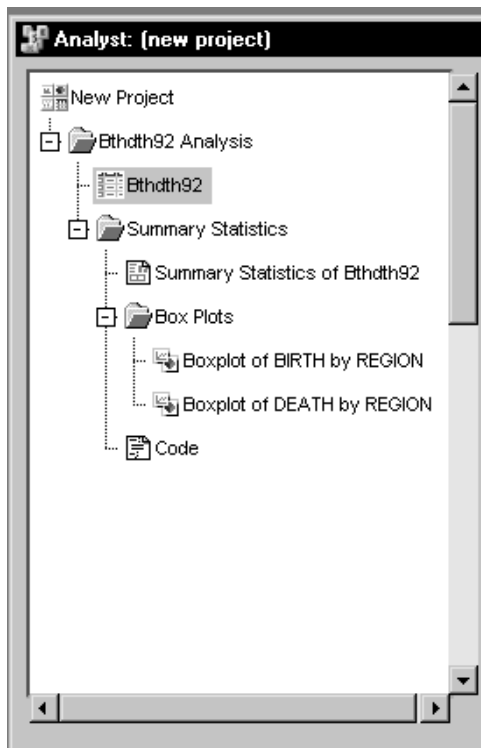
**Figure 7.8.** Summary Statistics: Plots Dialog

To perform the analysis, click **OK** in the main dialog.

### **Review the Results**

The results are presented in the project tree under the **Summary Statistics** folder, as displayed in [Figure 7.9](#). The four icons represent the summary statistics output, the box-and-whisker plots for each analysis variable, and the SAS programming statements (labeled **Code**) that generate the output.





**Figure 7.9.** Summary Statistics: Project Tree

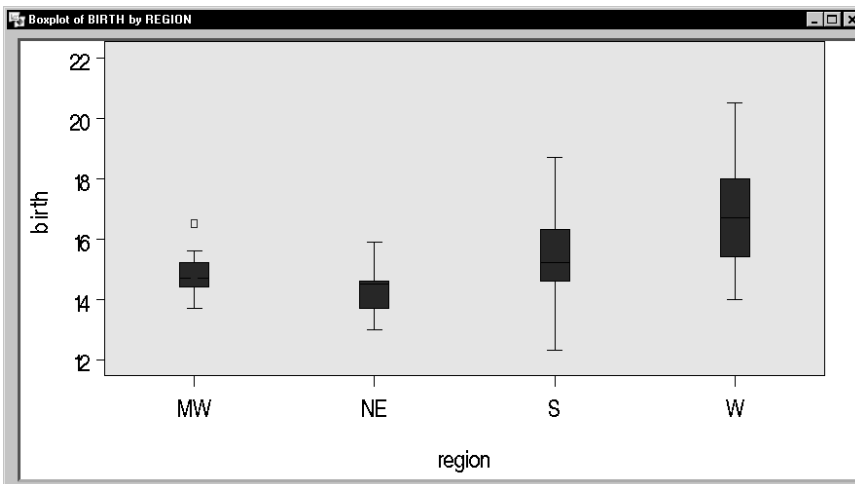
Double-click on any of the icons to display the corresponding information in a separate window.

Figure 7.10 displays, for each value of the classification variable **region**, the number of observations, the mean, the standard deviation, and the minimum and maximum values of each analysis variable. The western region has the highest birth rate (16.89) and the southern region has the highest death rate (10.15).

region	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
MW	12	birth	12	14.8250000	0.7581377	13.7000000	16.5000000
		death	12	8.5916667	1.0374833	7.1000000	10.2000000
NE	9	birth	9	14.3666667	0.8930286	13.0000000	15.9000000
		death	9	7.3777778	1.2194033	5.6000000	9.0000000
S	17	birth	17	15.4647059	1.4924565	12.3000000	18.7000000
		death	17	10.1529412	2.6241946	7.8000000	19.6000000
W	13	birth	13	16.8923077	2.1864970	14.0000000	20.5000000
		death	13	7.4769231	0.9670866	5.9000000	8.9000000

**Figure 7.10.** Summary Statistics: Statistics for birth and death

Figure 7.11 displays the box-and-whisker plot for the variable birth for each level of the region variable.



**Figure 7.11.** Summary Statistics: Box-and-Whisker Plot for Birth Rate by Region

This plot reveals a possible outlier in the birth rate for the midwestern region (region='MW'). The western region (region='W') is noticeable as the region with the highest birth rate.

## Examining the Distribution

You can examine the distributional properties of your data with the Distributions task. This task enables you to produce descriptive statistics for the variables, test the fit of several distributions to your data, and examine displays such as histograms and probability plots. In this task, interest lies in examining the birth and infant mortality rates for each region.

### Request a Distributions Analysis

To request the Distributions task, follow these steps:

1. Select **Statistics** → **Descriptive** → **Distributions . . .**
2. Select **birth** and **death** as the analysis variables.
3. Select **region** as the classification variable.

Figure 7.12 displays the Distributions main dialog with the preceding variable specifications.



Figure 7.12. Distributions Dialog

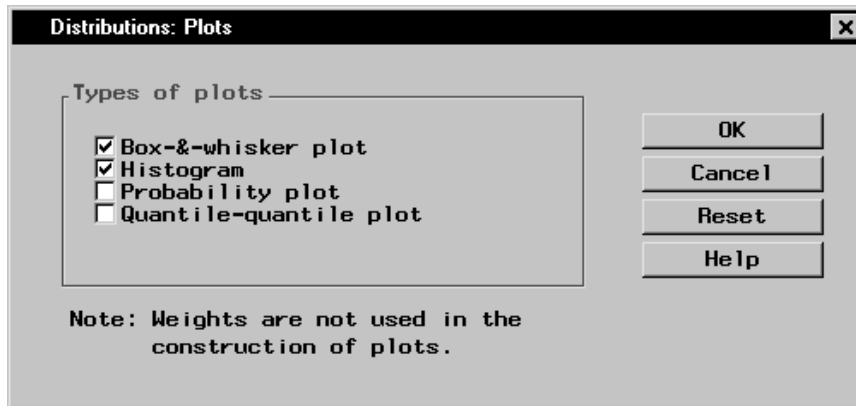
The default analysis provides moments, quartiles, and measures of variability.

### **Request Plots**

To request box-and-whisker plots and histograms, follow these steps:

1. Click on the **Plots** button.
2. Select **Box-&-whisker plot**.
3. Select **Histogram**.
4. Click **OK**.

Figure 7.13 displays the Plots dialog.



**Figure 7.13.** Distributions: Plots Dialog

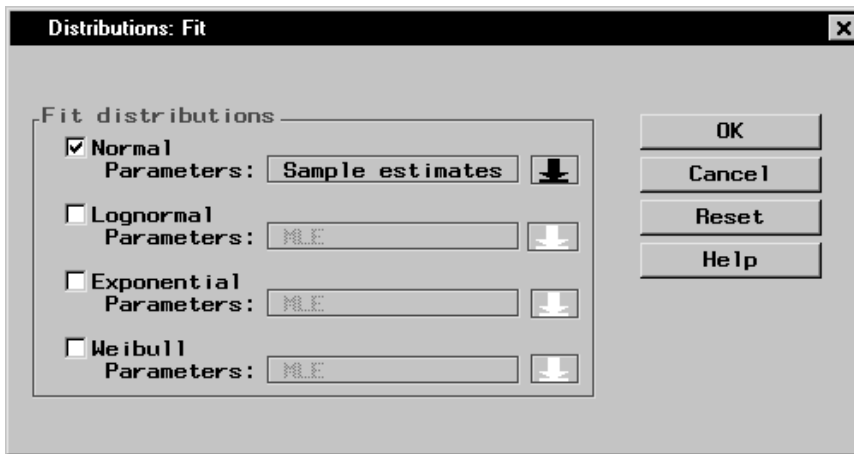
### **Request Fitted Distribution**

To fit a normal distribution to these data, follow these steps:

1. Click on the **Fit** button in the main dialog.
2. Select **Normal**.

By default, parameter values are calculated from the data when you fit the normal distribution. If you want to enter specific parameter values, click on the down arrow (displayed in [Figure 7.14](#)) and select **Enter values**. For the lognormal, exponential, and Weibull distributions, you can specify that parameters be calculated by maximum likelihood estimation (MLE), or you can enter specific parameter values.

3. Click **OK**.

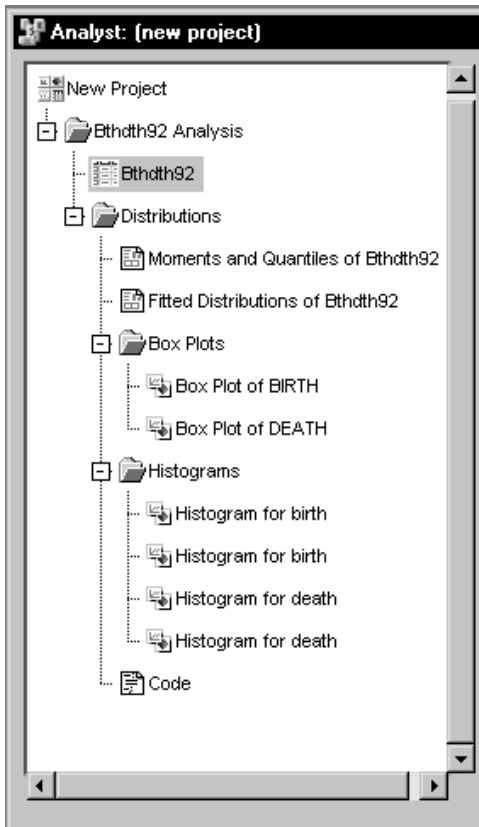


**Figure 7.14.** Distributions: Fit Dialog

When you have completed your selections, click **OK** in the main dialog to perform the analysis. The results are presented in the project tree displayed in [Figure 7.15](#).

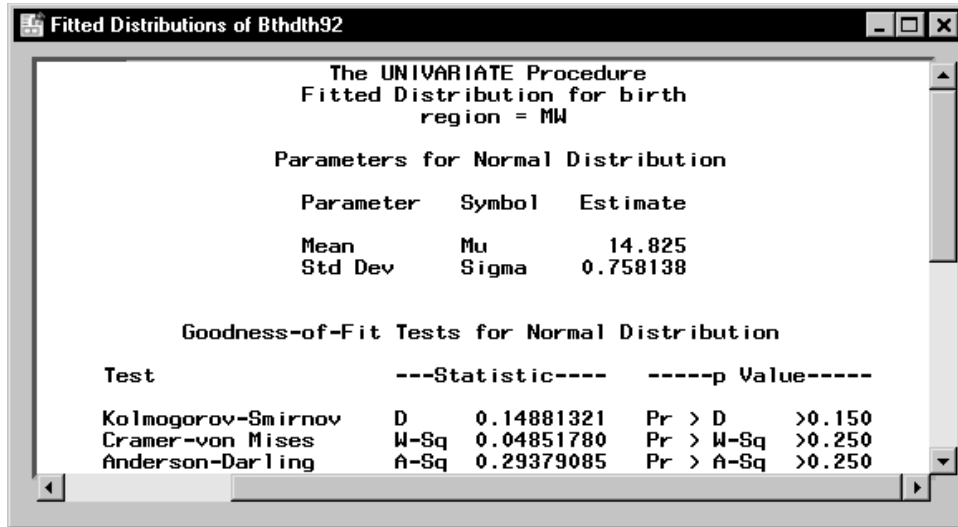
### **Review the Results**

Double-click on any of the resulting eight icons to display the corresponding output in a separate window.



**Figure 7.15.** Distributions: Project Tree

The Moments and Quantiles output provides summary information for each variable. [Figure 7.16](#) displays the output labeled Fitted Distributions of Bthdth92, which summarizes how closely the normal distribution fits each variable, by region.

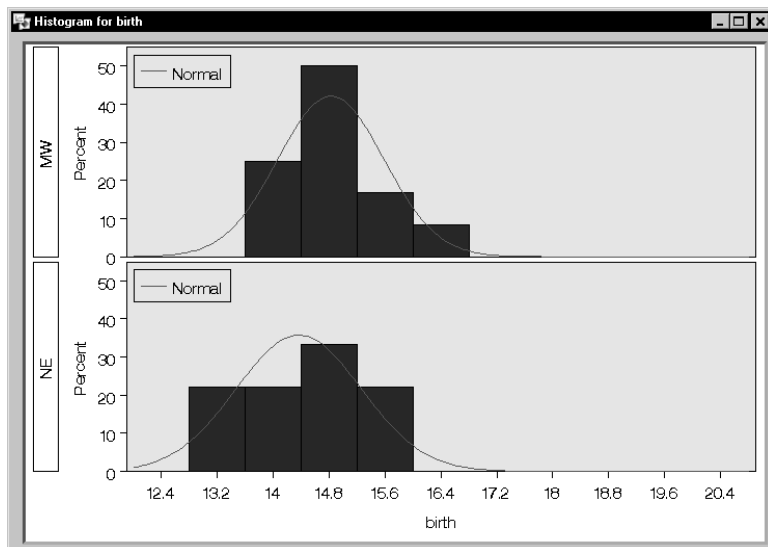


**Figure 7.16.** Distributions: Fitted Distributions Results

Based on the test results displayed in [Figure 7.16](#), the null hypothesis that the variable `birth` is normally distributed cannot be rejected at the  $\alpha = 0.05$  level of significance ( $p$ -values for all tests are greater than 0.15). The same is true for the variable `death` except for the southern region (`region='S'`). The hypothesis is rejected at the  $\alpha = 0.05$  level of significance for the death rate in the southern region.

Two sets of box plots and four sets of histograms are also produced. A single box-and-whisker plot is created for each of the two variables. The box-and-whisker plot for the variable `birth` is displayed when you double-click **Box Plot of BIRTH** in the project tree.

Two histograms are created for each variable. Each graphic contains a histogram for two levels of the classification variable `region`. The first histogram contains the information for the midwestern and northeastern regions (`region='MW'` and `region='NE'`), as displayed in [Figure 7.17](#). The second histogram (not shown) contains the information for the southern and western regions (`region='S'` and `region='W'`).



**Figure 7.17.** Distributions: Histogram for birth

The normal curve overlaid on the histogram displayed in [Figure 7.17](#) is the result of requesting a normal distribution fit in the Fit dialog ([Figure 7.14](#)). The statistical details of the fit are located in the output labeled Fitted Distributions of Bthdth92, which also includes the details of the fit for the variable death.

## Computing Correlations

You can use the Correlations task to compute pairwise correlation coefficients for the variables in your data set. The correlation is a measure of the strength of the linear relationship between two variables. This task can compute the standard Pearson product-moment correlations, nonparametric measures of association, partial correlations, and Cronbach's coefficient alpha. The task also can produce scatter plots with confidence ellipses.

The following example computes correlation coefficients for four variables in the Fitness data set. This data set contains measurements made on groups



of men taking a physical fitness course at North Carolina State University. The variables are as follows:

age	age, in years
weight	weight, in kilograms
oxygen	oxygen intake rate, in milliliters per kilogram of body weight per minute
runtime	time taken to run 1.5 miles, in minutes
rstpulse	heart rate while resting
runpulse	heart rate while running
maxpulse	maximum heart rate recorded while running
group	group number

This example includes looking at correlations between the variables `runtime`, `runpulse`, `maxpulse`, and `oxygen` and also producing the corresponding scatter plots with confidence ellipses.

### ***Open the Fitness Data Set***

To open the Fitness data set, follow these steps:

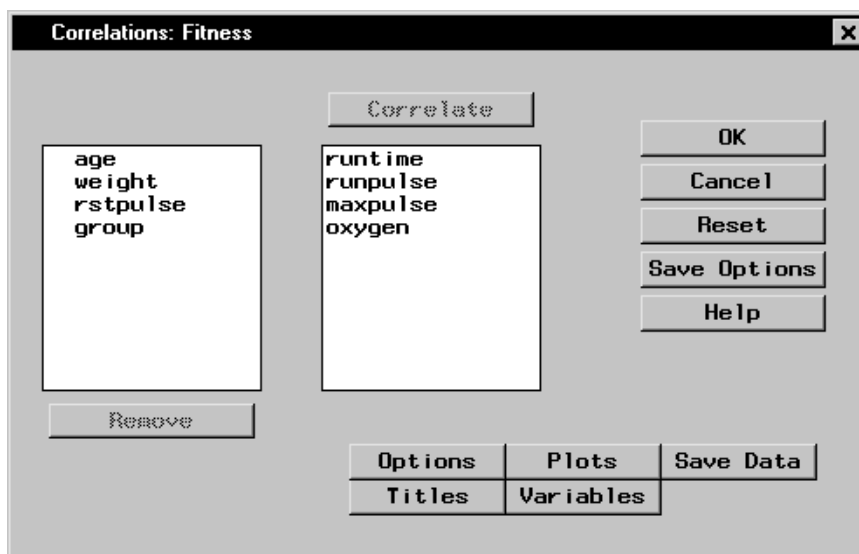
1. Select **Tools** → **Sample Data . . .**
2. Select **Fitness**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Fitness** from the list of members.
7. Click **OK** to bring the **Fitness** data set into the data table.

### Request Correlations

To compute correlations for variables in the Fitness data set, follow these steps:

1. Select **Statistics** → **Descriptive** → **Correlations . . .**
2. Select the variables runtime, runpulse, maxpulse, and oxygen to correlate.

Figure 7.18 displays the resulting Correlations dialog.



**Figure 7.18.** Correlations Dialog

If you click **OK** in the Correlations main dialog, the default output, which includes Pearson correlations, is produced. Or, you can request specific types of correlations by using the Options dialog.

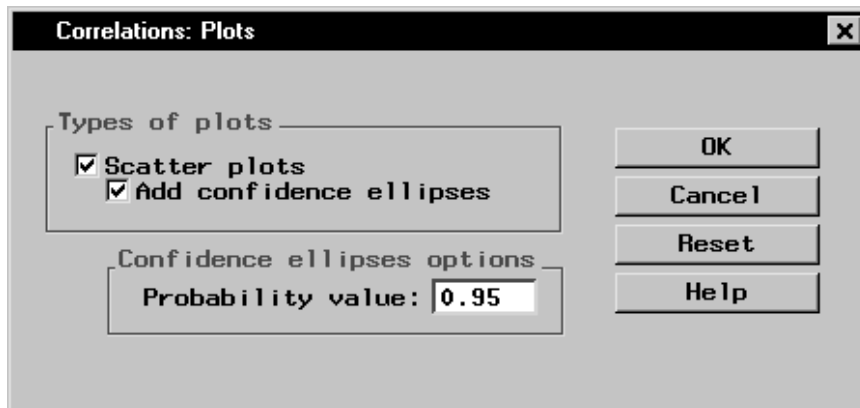
### Request a Scatter Plot

To request a scatter plot with a confidence ellipse, follow these steps:

1. Click on the **Plots** button.
2. Select **Scatter plots**.
3. Select **Add confidence ellipses**.

The confidence level used in calculating the confidence ellipse is 0.95. To use a different level, type that value in the **Probability value:** field, as displayed in [Figure 7.19](#).

4. Click **OK**.

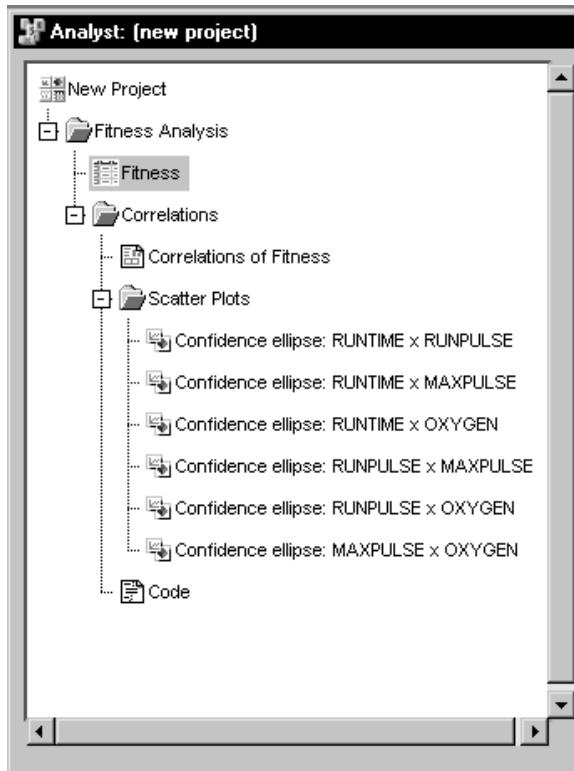


**Figure 7.19.** Correlations: Plots Dialog

Click **OK** in the main dialog to perform the analysis.

### Review the Results

The results are presented in the project tree, as displayed in [Figure 7.20](#).



**Figure 7.20.** Correlations: Project Tree

You can double-click on any of the resulting nodes in the project tree to view the information in a separate window.

Figure 7.21 displays univariate statistics for each of the analysis variables. The table provides the number of observations, the mean, the standard deviation, the sum, and the minimum and maximum values for each variable.

The screenshot shows a window titled "Correlations of Fitness" with the following content:

The CORR Procedure

4 Variables: runtime runpulse maxpulse oxygen

Simple Statistics

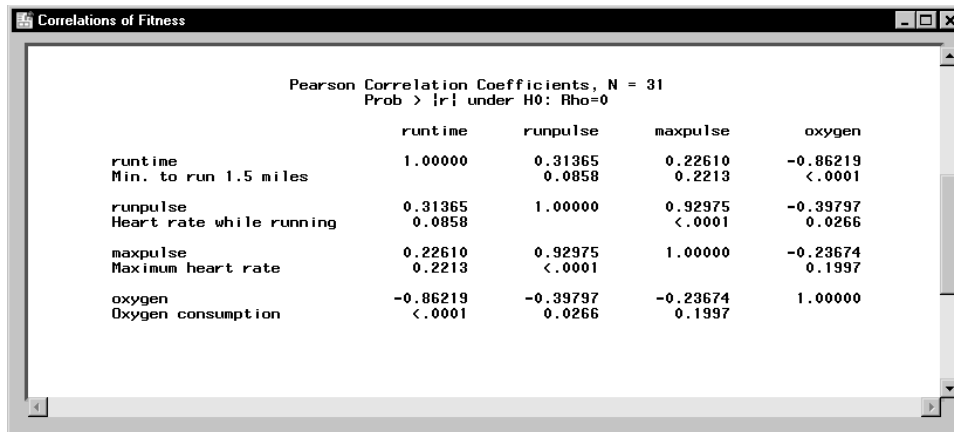
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
runtime	31	10.58613	1.38741	328.17000	8.17000	14.03000
runpulse	31	169.64516	10.25199	5259	146.00000	186.00000
maxpulse	31	173.77419	9.16410	5387	155.00000	192.00000
oxygen	31	47.37581	5.32723	1469	37.38800	60.05500

Simple Statistics

Variable	Label
runtime	Min. to run 1.5 miles
runpulse	Heart rate while running
maxpulse	Maximum heart rate
oxygen	Oxygen consumption

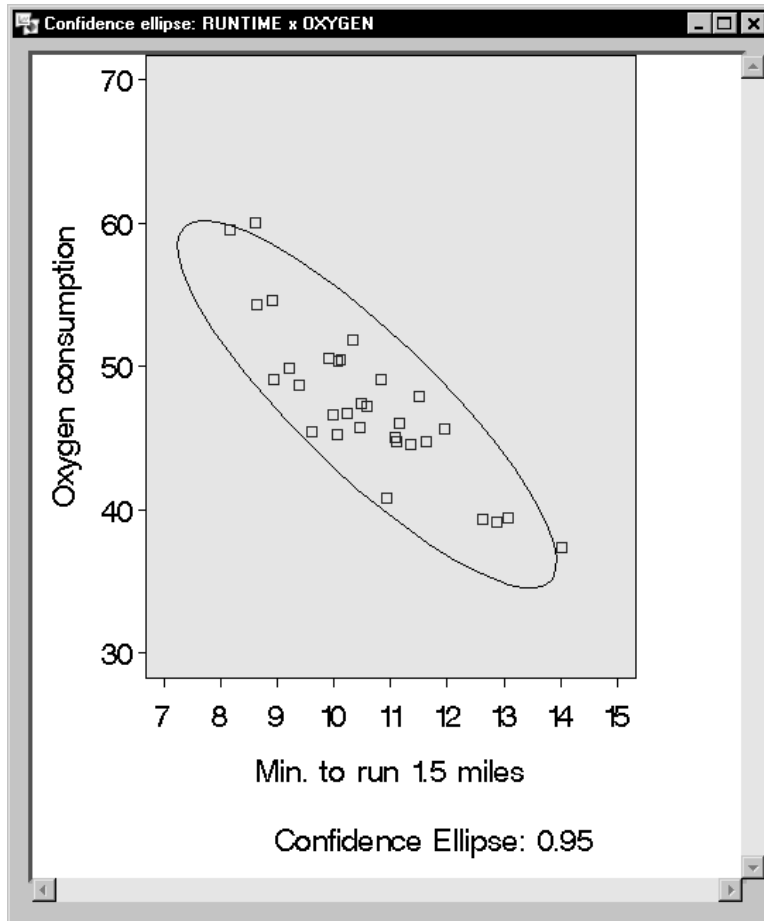
**Figure 7.21.** Correlations: Univariate Statistics

Figure 7.22 displays the table of correlations. The  $p$ -value, which is the significance probability of the correlation, is displayed under each of the correlation coefficients. For example, the correlation between the variables `maxpulse` and `runtime` is 0.22610, with an associated  $p$ -value of 0.2213, and the correlation between the variables `oxygen` and `runpulse` is  $-0.39797$ , with an associated  $p$ -value of 0.0266.



**Figure 7.22.** Correlations: Table of Correlations

Six scatter plots, each of which includes a 95% confidence ellipse, are produced in this analysis. Each plot displays the relationship between one pair of the analysis variables. The scatter plot of runtime versus oxygen is displayed in [Figure 7.23](#).



**Figure 7.23.** Correlations: Scatter Plot with Confidence Ellipse

Confidence ellipses are used as a graphical indicator of correlation. When two variables are uncorrelated, the confidence ellipse is circular in shape. The ellipse becomes more elongated the stronger the correlation is between two variables.

---

## References

- SAS Institute Inc. (2000), *SAS Procedures Guide, Version 8*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.
- Schlotzhauer, Sandra D. and Littell, Ramon C. (1991), *SAS System for Elementary Statistical Analysis, Second Edition*, Cary, NC: SAS Institute Inc.
- U.S. Bureau of the Census (1995), *Statistical Abstract of the United States*, Washington, D.C.



# Chapter 8

## Hypothesis Tests

### Chapter Contents

---

<b>Introduction</b> . . . . .	211
<b>One-Sample t-Test</b> . . . . .	212
<b>Paired t-test</b> . . . . .	218
<b>Two-Sample Test for Proportions</b> . . . . .	224
<b>Two-Sample Test for Variances</b> . . . . .	227
<b>Discussion of Other Tests</b> . . . . .	231
<b>References</b> . . . . .	233



# Chapter 8

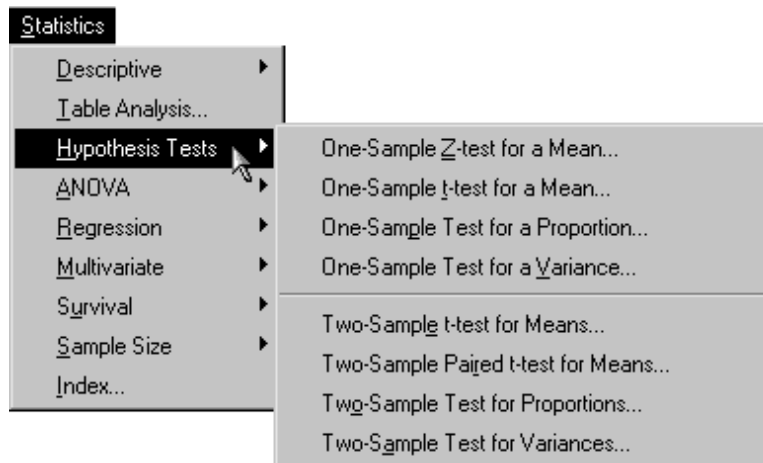
## Hypothesis Tests

---

### Introduction

Hypothesis tests are frequently performed for one and two samples. For one sample, you are often interested in whether a population characteristic such as the mean is equivalent to a certain value. For two samples, you may be interested in whether the true means are different. When you have paired data, you may be interested in whether the mean difference is zero.

Statistical hypothesis tests depend on a statistic designed to measure the degree of evidence for various alternative hypotheses. You compute the value of the statistic for your sample. If the value is improbable under the hypothesis you want to test, then you reject the hypothesis.



**Figure 8.1.** Hypothesis Tests Menu

The Analyst Application enables you to perform hypothesis tests for means, proportions, and variances for either one or two samples.

The examples in this chapter demonstrate how you can use the Analyst Application to perform a one-sample  $t$ -test, a paired  $t$ -test, a two-sample test for proportions, and a two-sample test for variances. Additionally, the section “[Discussion of Other Tests](#)” on page 231 provides information on other hypothesis tests you can perform with the Analyst Application.

---

## One-Sample $t$ -Test

The One-Sample  $t$ -Test task enables you to test whether the mean of a variable is less than, greater than, or equal to a specific value. The observed mean of the variable is compared to this value.

The data set analyzed in the following example, `Bthdth92`, is taken from the 1995 Statistical Abstract of the United States, and it contains measures of the birth rate and infant mortality rate for 1992 in the United States. Information is provided for the 50 states and the District of Columbia, grouped by region.

Suppose you want to determine whether the average infant mortality rate in the United States is equal to a specific value. Note that the one-sample  $t$ -test is appropriate in this situation because the standard deviation of the population from which the data arise is unknown. When you know the standard deviation of the population, use the One-Sample Z-Test for a Mean task (see the section “[Discussion of Other Tests](#)” on page 231 for more information).

### Open the `Bthdth92` Data Set

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select `Bthdth92`.
3. Click **OK** to create the sample data set in your `Sasuser` directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select `Sasuser` from the list of **Libraries**.
6. Select `Bthdth92` from the list of members.
7. Click **OK** to bring the `Bthdth92` data set into the data table.

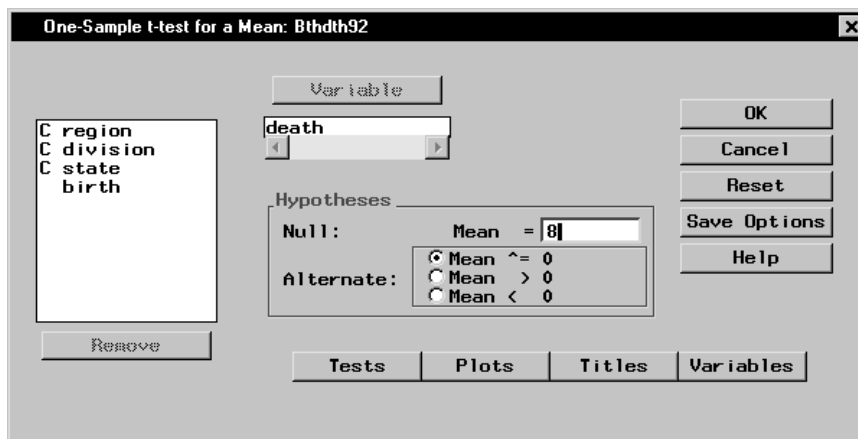
### Request a One-Sample *t*-Test

To test whether the average infant mortality rate is equal to 8, follow these steps:

1. Select **Statistics** → **Hypothesis Tests** → **One-Sample *t*-Test for a Mean . . .**
2. Select **death** as the variable to be analyzed.
3. Enter **8** in the box labeled **Null: Mean =** and press **Enter**.

Your alternative hypothesis can be that the mean is less than, greater than, or not equal to a specified value. In this example, the alternative hypothesis is that the mean of the variable **death** is not equal to 8.

In [Figure 8.2](#), the one-sample *t*-test dialog defines the null and alternative hypotheses and specifies **death** as the variable to be tested.



**Figure 8.2.** One-Sample *t*-Test Dialog

The default one-sample *t*-test task includes sample statistics for the variable **death** and the hypothesis test results.

### Compute a Confidence Interval for the Mean

To produce a confidence interval for the mean in addition to the hypothesis test, follow these steps:

1. Click on the **Tests** button in the main dialog.
2. Select **Interval** to request a two-sided confidence interval for the mean.

You can choose either a one-sided or a two-sided confidence interval for the mean. The selections **Lower bound** and **Upper bound** specify one-sided confidence bounds.

The default confidence level is 95%. You can click on the down arrow to select another confidence level, or you can enter a confidence level in the box.

3. Click **OK** to return to the main dialog.

Figure 8.3 displays the selection of a 95% two-sided confidence interval for the mean. Note that you can also request a retrospective power analysis of the test in the **Power Analysis** tab.

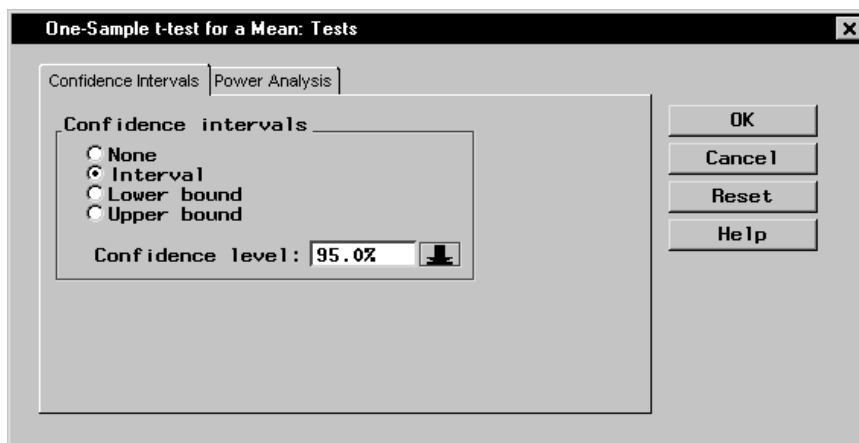


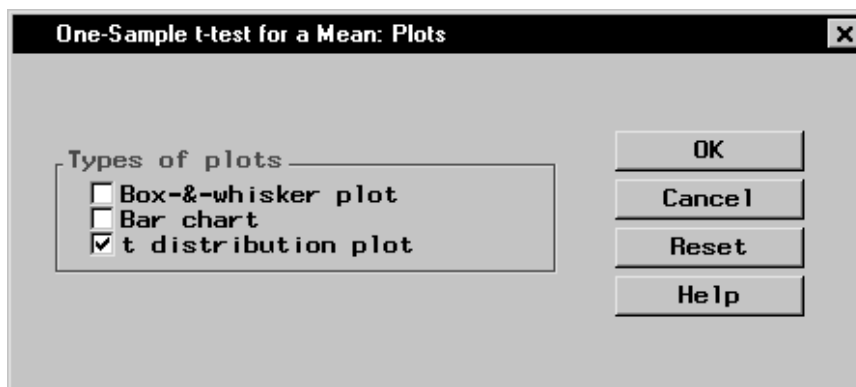
Figure 8.3. One-Sample t-Test: Tests Dialog

### Request a *t* Distribution Plot

To request a *t* distribution plot in addition to the hypothesis test, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **t distribution plot**.
3. Click **OK** to return to the main dialog.

Figure 8.4 displays the Plots dialog with **t distribution plot** selected.



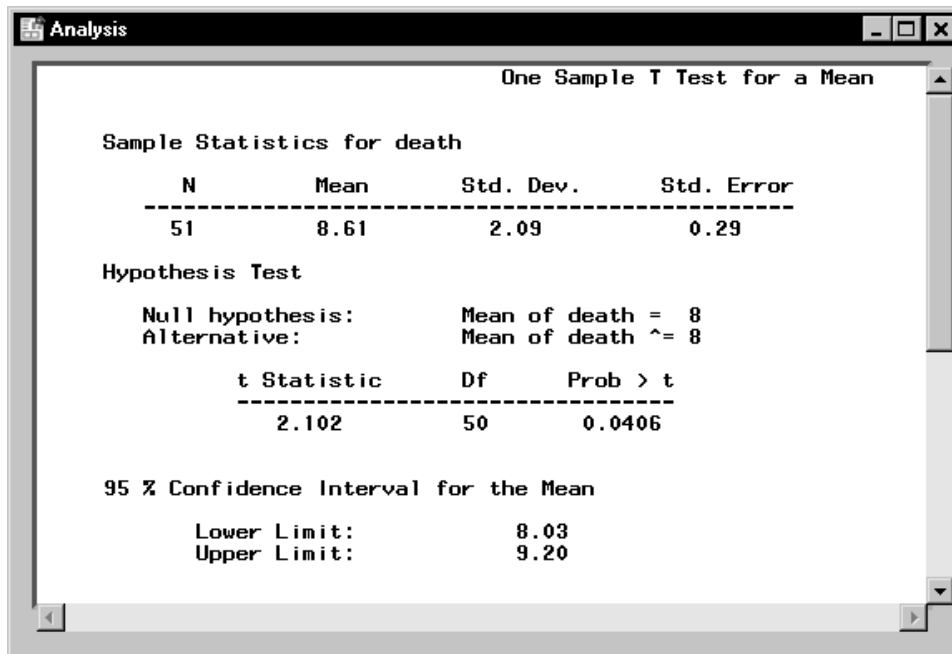
**Figure 8.4.** One-Sample t-Test: Plots Dialog

Click **OK** in the main dialog to perform the analysis.

### Review the Results

The results of the hypothesis test are displayed in Figure 8.5. The output includes the “Sample Statistics” table for the variable `death`, the hypothesis test results, and the 95% confidence interval for the mean.

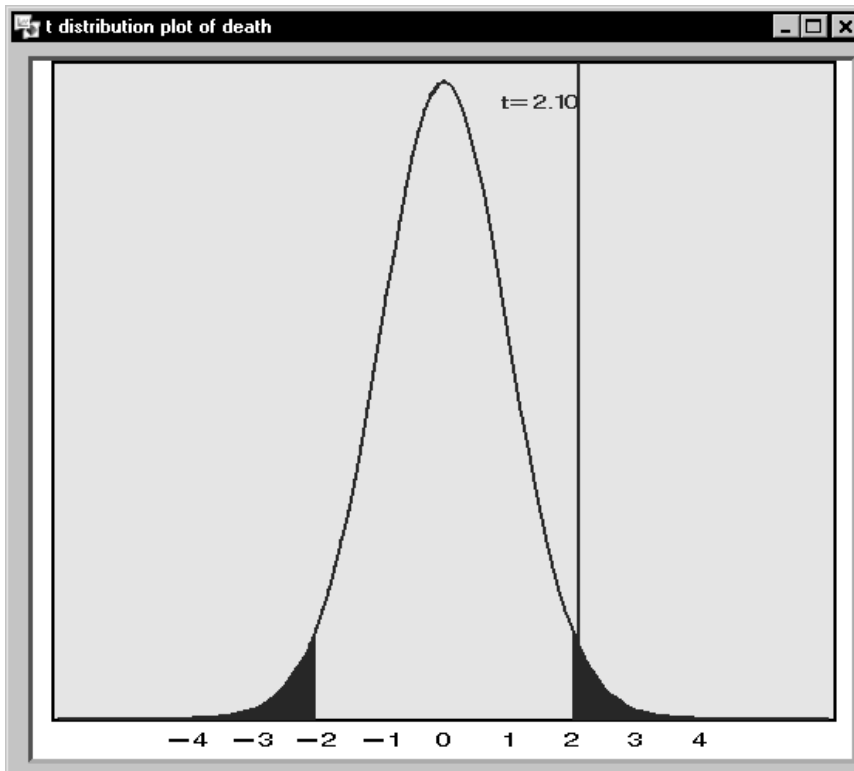
The mean of the variable `death` is 8.61, which is greater than the specified test value of 8.



**Figure 8.5.** One-Sample t-Test: Output

The  $t$  statistic of 2.102 and the associated  $p$ -value (0.0406) provide evidence at the  $\alpha = 0.05$  level that the average infant mortality rate is not equal to 8. The confidence interval indicates that you can be 95% confident that the true mean of the variable lies within the interval [8.03, 9.20].





**Figure 8.6.** One-Sample  $t$ -Test:  $t$  Distribution Plot

The requested  $t$  distribution plot is displayed in Figure 8.6. The plot depicts the calculated  $t$  statistic superimposed on a  $t$  distribution density function with 50 degrees of freedom.

Because this analysis requests a two-tailed test, two critical regions are shaded, one in each of the left and right tails. The alpha level for the test is 0.05; thus, each region represents 2.5% of the area under the curve. In a one-tailed test at the  $\alpha = 0.05$  level, the critical region appears in one tail only, and it represents 5% of the area under the curve.

Here, the  $t$  statistic falls in the shaded region. Thus, the null hypothesis is rejected.

---

## Paired t-test

The Paired  $t$ -test enables you to determine whether the means of paired samples are equal. The term *paired* means that there is a correspondence between observations from each population. For example, the birth and death data analyzed in the preceding section are considered to be paired data because, in each observation, the variables **birth** and **death** correspond to the same state.

Suppose that you want to determine whether the means for the birth rate and the infant mortality rate are equal. Analyst provides the Two-Sample Paired  $t$ -test for Means task, which tests the equality of means of two paired samples. The two samples in this example are the birth rate (**birth**) and the infant mortality rate (**death**) for each state.

### Open the Bthdth92 Data Set

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Bthdth92.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Bthdth92 from the list of members.
7. Click **OK** to bring the Bthdth92 data set into the data table.

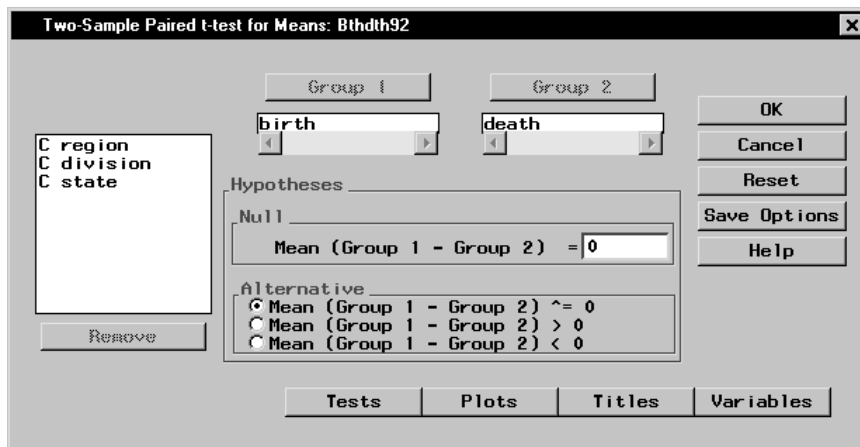
### Request a Paired t-Test

To perform this analysis, follow these steps:

1. Select **Statistics** → **Hypothesis Tests** →  
**Two-Sample Paired t-test for Means** . . .
2. Select the variable **birth** as the Group 1 variable.
3. Select the variable **death** as the Group 2 variable.

The test of interest is whether the difference of the means is zero. This is the default value in Analyst, although you can specify other values as well.

You can choose one of three alternative hypotheses. The default is that the difference between the means is not equal to the specified difference, which is the two-sided alternative. The one-sided alternatives are that the difference is greater than, or less than, the difference specified in the null hypothesis.



**Figure 8.7.** Paired t-test Dialog

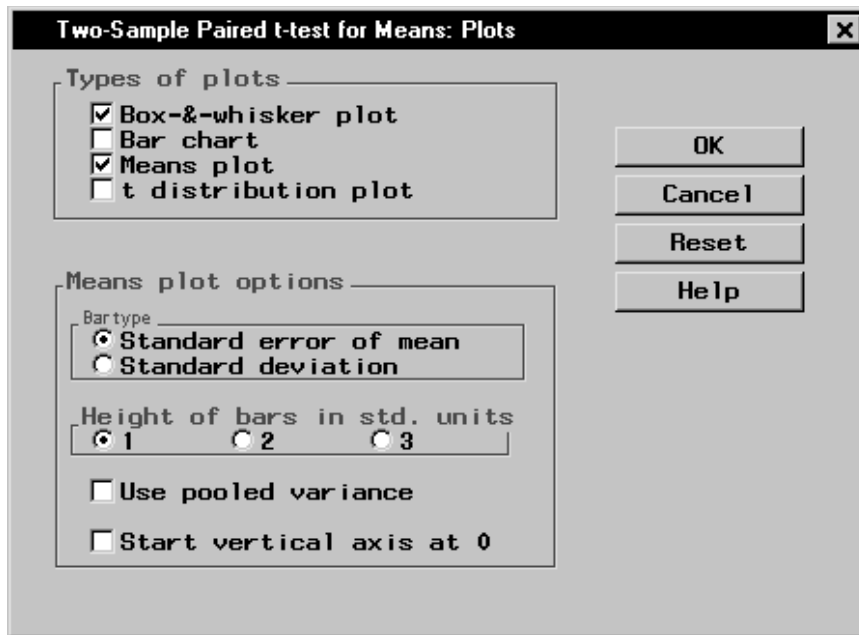
In [Figure 8.7](#), the null hypothesis specifies that the means of the variables birth and death are equal (or, equivalently, that the difference between the means is 0). The alternative hypothesis is that the two means are not equal.

### **Request Plots**

To specify a box-and-whisker plot and a means plot in addition to the hypothesis test, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Box-&-whisker plot**.
3. Select **Means plot**.
4. Click **OK**.

Figure 8.8 displays the Plots dialog with **Box-&-whisker plot** and **Means plot** selected.

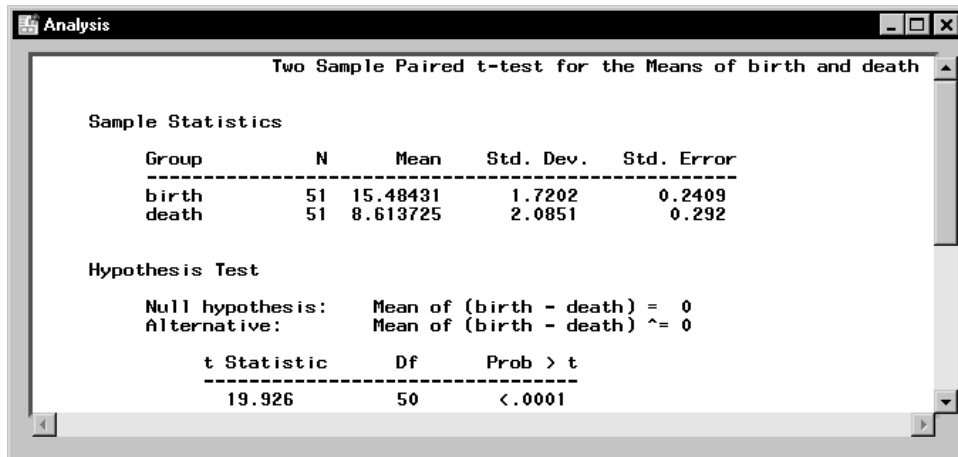


**Figure 8.8.** Paired t-test: Plots Dialog

Click **OK** in the main dialog to perform the analysis.

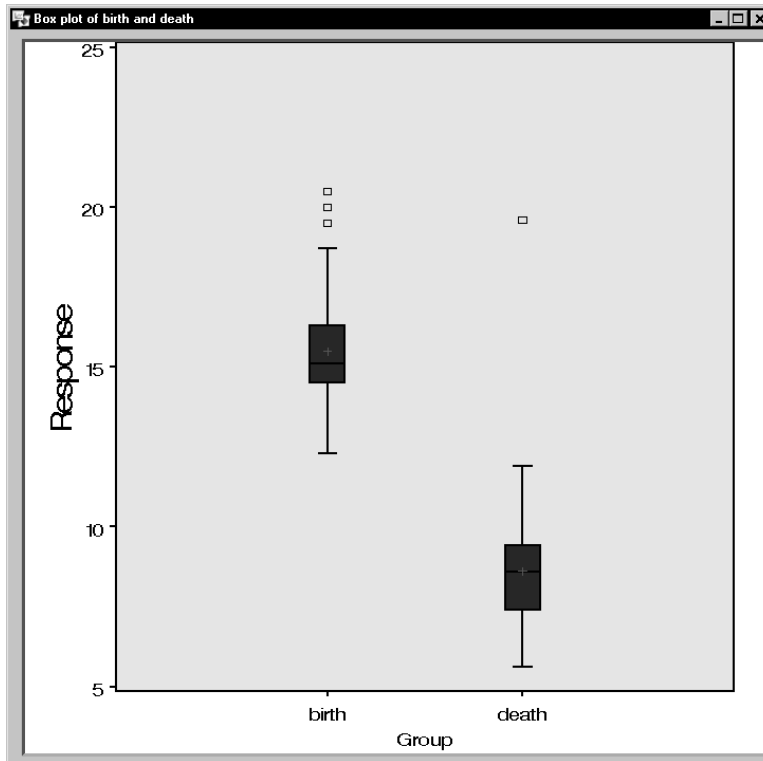
### **Review the Results**

The results of the analysis, displayed in Figure 8.9, contain the mean, standard deviation, and standard error of the mean for both variables. The “Hypothesis Test” table provides the observed  $t$  statistic, the degrees of freedom, and the associated  $p$ -value of the test.



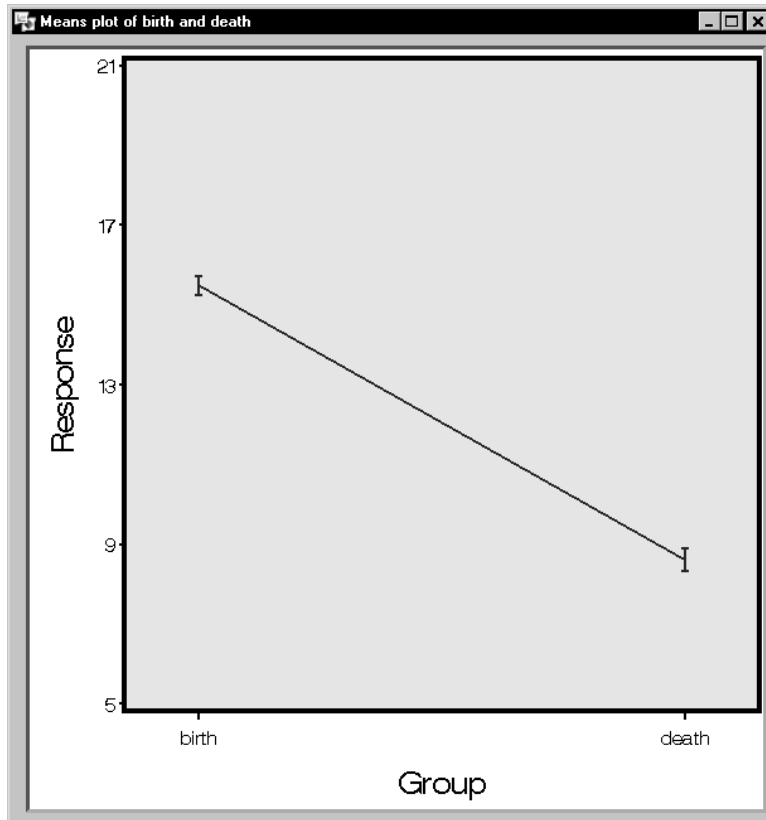
**Figure 8.9.** Paired t-test: Results

In [Figure 8.9](#), the “Sample Statistics” table shows that the mean of the variable `birth` is larger than that of the variable `death`. In the “Hypothesis Test” table, the  $t$  statistic (19.926) and associated  $p$ -value ( $< 0.0001$ ) indicate that the difference between the two means is statistically very significant.



**Figure 8.10.** Paired t-test: Box-and-Whisker Plot

Figure 8.10 displays the side-by-side box plots of birth and death. Observations that fall beyond the whiskers are individually identified with a square symbol.



**Figure 8.11.** Paired t-test: Means Plot

The means and standard error plot displayed in [Figure 8.11](#) provides another view of the two variables. The means plot depicts an interval centered on the sample mean for each variable. The vertical line interval extends two standard deviations on either side of the mean.

---

## Two-Sample Test for Proportions

In the Two-Sample Test for Proportions task, you can determine whether two probabilities are the same.

The data analyzed in this example are taken from a study measuring the accuracy of two computer programs. Each program searches the World Wide Web and returns a list of web pages that meet a particular set of specified criteria. The data set **Search** contains two samples in which each observation is either 'yes' or 'no'. A response of 'yes' indicates that the program returns the desired page at the top of the list of potential pages; a value of 'no' indicates that this is not the case. The data set contains the results of 535 searches using an older search program and 409 searches using a new program. The variables containing the results for the old and new programs are named **oldfind** and **newfind**, respectively.

Suppose that you want to determine whether the probability of a correct search by the new algorithm is higher than that for the old algorithm. That is, you want to determine whether you can reject the null hypothesis that the two probabilities are equal in favor of the alternative that the new probability is larger. The values for analysis are contained in the two variables **oldfind** and **newfind**.

### **Open the Search Data Set**

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select **Search**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Search** from the list of members.
7. Click **OK** to bring the **Search** data set into the data table.



### Request a Two-Sample Test for Proportions

To perform the analysis, follow these steps:

1. Select **Statistics** → **Hypothesis Tests** → **Two-Sample Test for Proportions . . .**
2. Select **Two variables** in the box labeled **Groups are in**.
3. Select the variable **newfind** as the Group 1 variable.
4. Select the variable **oldfind** as the Group 2 variable.
5. Select the **Level of Interest** by clicking on the down arrow and selecting **yes** to test whether the two groups have the same proportions of success.
6. Specify the **Alternative** hypothesis by selecting **Prop 1 - Prop 2 > 0**.

Note that, if your data are arranged so that the values for the two groups are contained in a single variable, you can define the dependent and group variables by selecting **One variable** in the box labeled **Groups are in**.

Figure 8.12 displays the Two-Sample Test for Proportions dialog.

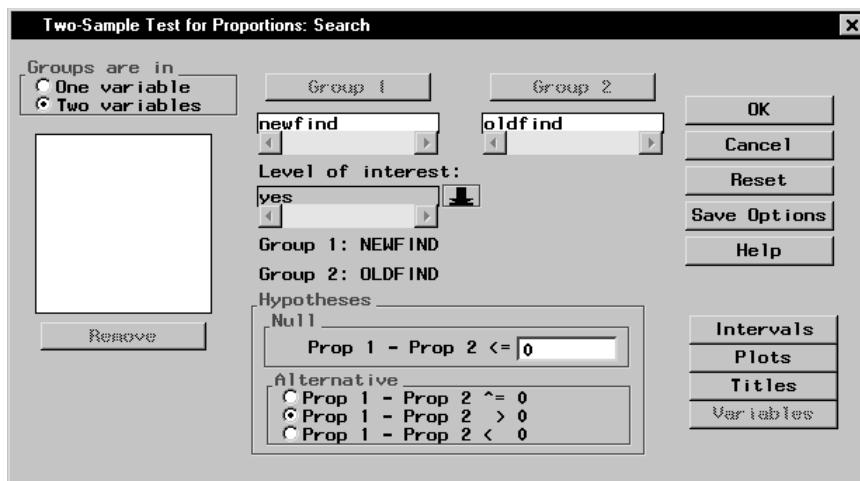


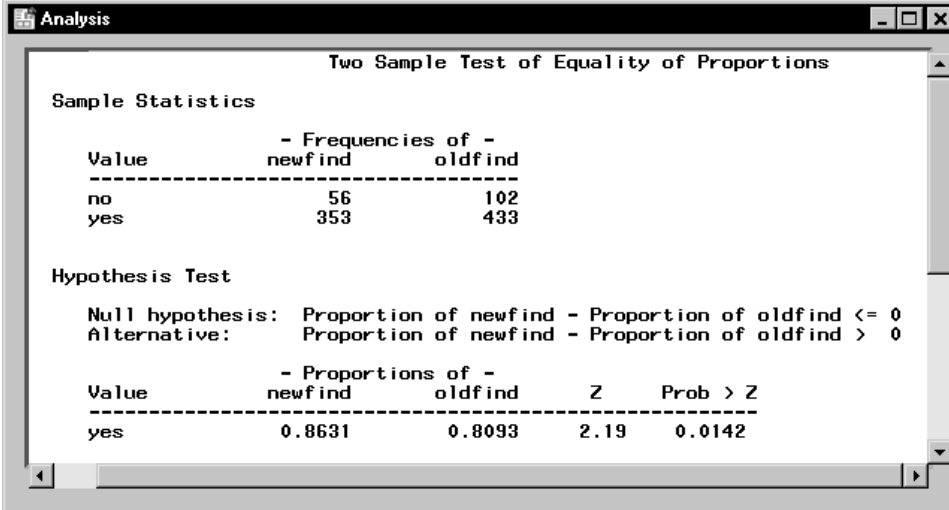
Figure 8.12. Two-Sample Test for Proportions Dialog

In Figure 8.12, the null hypothesis specifies that the proportions of success for the algorithms are equal (or, equivalently, that the difference between the proportions is 0). The alternative hypothesis is that the probability of a correct search by the new algorithm is higher than that for the old algorithm.

Click **OK** in the main dialog to perform the analysis.

### Review the Results

The results of the hypothesis test are displayed in Figure 8.13.



The screenshot shows a window titled "Analysis" with a subtitle "Two Sample Test of Equality of Proportions". It displays the following data:

Two Sample Test of Equality of Proportions				
<b>Sample Statistics</b>				
	- Frequencies of -			
Value	newfind	oldfind		
no	56	102		
yes	353	433		
<b>Hypothesis Test</b>				
Null hypothesis: Proportion of newfind - Proportion of oldfind <= 0				
Alternative: Proportion of newfind - Proportion of oldfind > 0				
	- Proportions of -			
Value	newfind	oldfind	Z	Prob > Z
yes	0.8631	0.8093	2.19	0.0142

**Figure 8.13.** Two-Sample Test for Proportions: Results

The “Sample Statistics” table lists the frequency of ‘yes’ and ‘no’ responses for each variable. The “Hypothesis Test” table displays the null and alternative hypotheses and the results of the test.

The observed proportion of ‘yes’ responses is 0.8631 for the *newfind* variable, and 0.8093 for the *oldfind* variable. The *Z* statistic of 2.19 and associated *p*-value of 0.0142 indicate that the proportion of successful searches is significantly larger for the new search algorithm.

---

## Two-Sample Test for Variances

In the Two-Sample Test for Variances task, you can test whether two variables have different variances, or, if you have a single variable that contains values for two groups, you can determine whether the variance differs between the groups.

The data set analyzed in this example, **Gpa**, contains test scores for 224 students. The data include the students' grade point averages (the variable **gpa**), high school scores in mathematics, science, and English (the variables **hsm**, **hss**, and **hse**, respectively), and SAT math and verbal scores (the variables **satm** and **satv**, respectively).

Suppose that you want to examine the difference in grade point averages between males and females. You can use the two-sample test for variances to test whether the variance of the grade point average differs between males and females.

### **Open the Gpa Data Set**

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select **GPA**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Gpa** from the list of members.
7. Click **OK** to bring the **Gpa** data set into the data table.

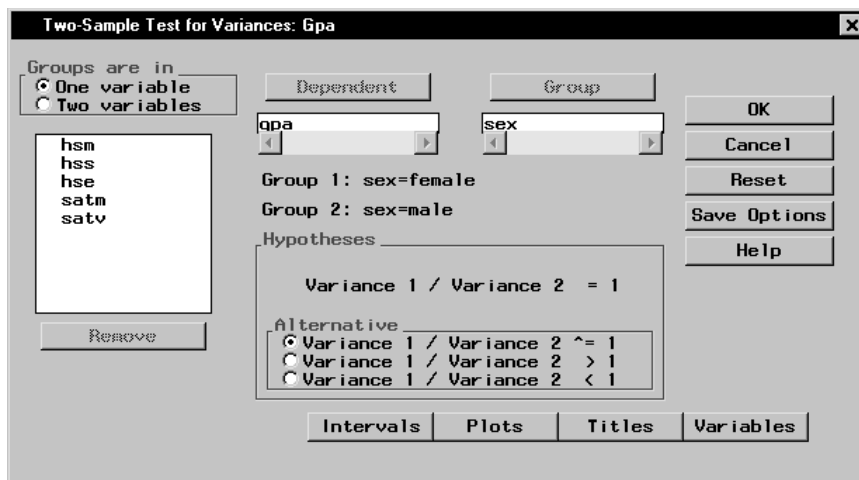
### **Request a Two-Sample Test for Variances**

To perform the hypothesis test, follow these steps:

1. Select **Statistics** → **Hypothesis Tests** → **Two-Sample Test for Variances . . .**
2. Ensure that **One variable** is selected in the box labeled **Groups are in**.
3. Select the variable **gpa** as the Dependent variable.
4. Select the variable **sex** as the Group variable.

If your data are arranged so that the values for both groups are contained in two variables, you can define the two groups by checking the **Two variables** selection in the box labeled **Groups are in**.

The null hypothesis for the test is that the two variances are equal (or, equivalently, that their ratio is equal to 1). You can specify the type of alternative hypothesis. The three choices are that Variance 1 is not equal to, is greater than, or is less than Variance 2. In [Figure 8.14](#), the alternative hypothesis states that the two variances are not equal, which is the two-sided alternative hypothesis.



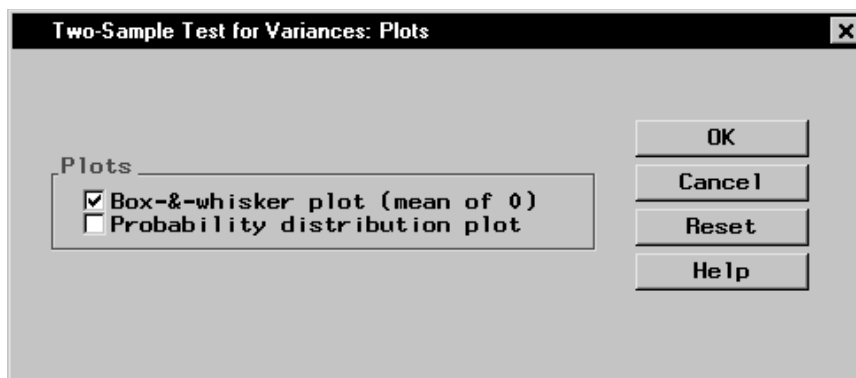
**Figure 8.14.** Two-Sample Test for Variances Dialog

### Request a Box-&-Whisker Plot

To request a box-and-whisker plot in addition to the hypothesis test, follow these steps:

1. Click on the **Plots** button.
2. Select **Box-&-whisker plot**.
3. Click **OK**.

Figure 8.15 displays the Plots dialog with **Box-&-whisker plot** selected. Note that the plot is constructed to have a mean of zero.



**Figure 8.15.** Two-Sample Test for Variances: Plots Dialog

Click **OK** in the Two-Sample Test for Variances dialog to perform the hypothesis test.

### Review the Results

Figure 8.16 displays the results of the hypothesis test. The output contains the results of the hypothesis test, including summary statistics, the  $F$  statistic, and the associated  $p$ -value.

Two Sample Test for Variances of gpa within sex

Sample Statistics

sex Group	N	Mean	Std. Dev.	Variance
female	145	4.607724	0.8068	0.650883
male	79	4.685696	0.7288	0.531086

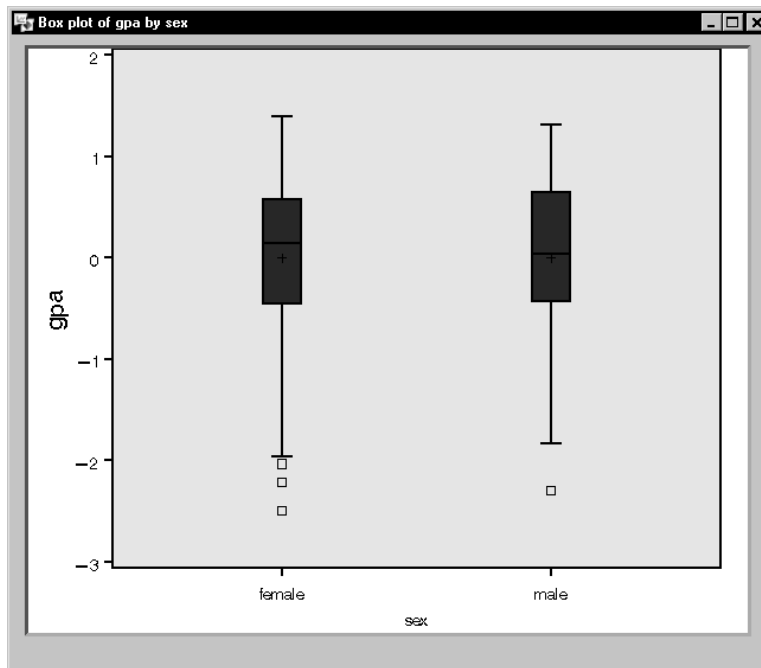
Hypothesis Test

Null hypothesis: Variance 1 / Variance 2 = 1  
 Alternative: Variance 1 / Variance 2  $\neq$  1

F	- Degrees of Freedom -		Pr > F
	Numer.	Denom.	
1.23	144	78	0.3222

**Figure 8.16.** Two-Sample Test for Variances: Output

The “Sample Statistics” table displays the variance of the variable `gpa` for both females (0.6509) and males (0.5311). The “Hypothesis Test” table displays the test statistics: the  $F$  value is 1.23 and the resulting  $p$ -value is 0.3222. Thus, the data give no evidence for rejecting the hypothesis of equal variances.



**Figure 8.17.** Two-Sample Test for Variances: Box-and-whisker Plot

Figure 8.17 displays the box-and-whisker plot. Observations that fall beyond the whiskers are identified with a square symbol.

The box-and-whisker plot displays the amount of spread and the range for the two variables. The two groups do not appear to be appreciably different.

---

## Discussion of Other Tests

The following descriptions provide an overview of other hypothesis tests available in the Analyst Application.

### ***One-Sample Z-Test for a Mean***

In the One-Sample Z-Test for a Mean task, you can test whether the mean of a population is equal to the value you specify in the null hypothesis. This test is appropriate when the population standard deviation or variance is known, and your data are either normally distributed or you have a large number

of observations. Generally, a sample size of at least 30 is considered to be sufficient.

The default output from the test includes summary statistics for the selected variable, the  $Z$  statistic, and the associated  $p$ -value.

### **One-Sample Test for a Proportion**

In the One-Sample Test for a Proportion task, you can test whether the proportion of a population giving a certain response is equal to the proportion you specify in the null hypothesis.

The default output from this test provides a frequency table of responses versus the analysis variable, the observed proportion, the  $Z$  statistic, and the associated  $p$ -value.

### **One-Sample Test for a Variance**

In the One-Sample Test for a Variance task, you can test whether the variance of a population is equal to the value you specify in the null hypothesis.

The default output from this test includes summary statistics for the selected variable, the chi-square statistic, and the associated  $p$ -value.

### **Two-Sample $t$ -Test for Means**

In the Two-Sample  $t$ -Test for Means task, you can test whether the means of two populations are equal or, optionally, whether they differ by a specified amount. Two-sample data arise when two independent samples are observed, possibly with different sample sizes. Note that, if the two samples are not independent, the two-sample  $t$ -test is inappropriate and you should use instead the Two-Sample Paired  $t$ -Test for Means task (see the section “Paired  $t$ -test” beginning on page 218 for more information).

The default output from the test includes summary statistics for the two samples, two  $t$  statistics, and the associated  $p$ -values. The first  $t$  statistic assumes the population variances of the two groups are equal; the second statistic is an approximate  $t$  statistic and should be used when the population variances of the two groups are potentially unequal.



---

## References

- SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.
- Schlotzhauer, Sandra D. and Littell, Ramon C. (1991), *SAS System for Elementary Statistical Analysis, Second Edition*, Cary, NC: SAS Institute Inc.
- U.S. Bureau of the Census (1995), *Statistical Abstract of the United States*, Washington, D.C.



# Chapter 9

## Table Analysis

### Chapter Contents

---

<b>Introduction</b> . . . . .	237
<b>Association in a <math>2 \times 2</math> Table</b> . . . . .	238
<b>Exact Test</b> . . . . .	245
<b>Association in Sets of Tables</b> . . . . .	251
<b>Observer Agreement</b> . . . . .	261
<b>References</b> . . . . .	266



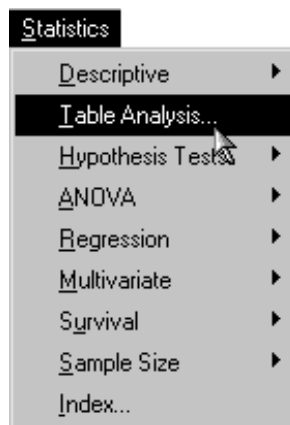
# Chapter 9

## Table Analysis

---

### Introduction

Often you need to analyze the information in a table, sometimes called a contingency table or a crossclassification table. You may analyze a single table, or you may analyze a set of tables. You are also often concerned with evaluating the presence of *association* in a table, or whether there is some sort of relationship between the variable determining the rows of the table and the variable determining the columns of the table. If there is an inherent ordering in the rows or columns of the table, the association may be linear. Various chi-square statistics such as the Pearson chi-square and the likelihood ratio chi-square are used to assess association.



**Figure 9.1.** Table Analysis Selection Menu

Besides assessing the presence of association, you may also be interested in computing a *measure of association*, or a statistic that provides some understanding of the strength of the association. The odds ratio is a standard measure of association often used in medical and epidemiological studies.

Using the Table Analysis task, not only can you analyze a single table, but you can also analyze sets of tables. This provides a way to control, or adjust for, a covariate, while assessing association of the rows and columns of the tables. Extended Mantel-Haenszel statistics, also called Cochran-Mantel-Haenszel statistics, provide a way to utilize all the information in the constituent tables in a test for the hypothesis of association. Tables may also contain information from observer agreement studies in which the evaluations or assessments of two different observers are collected. Statistics called measures of agreement assess how closely the observers agree.

The Table Analysis task provides chi-square tests of association for the  $r \times c$  table, including statistics such as the Pearson chi-square and likelihood ratio test, and it also computes extended Mantel-Haenszel tests for sets of tables. Fisher's exact test can be computed for both the  $2 \times 2$  and  $r \times c$  table. In addition, the Table Analysis task also provides measures of association such as the odds ratio and relative risk for the  $2 \times 2$  table as well as gamma, tau- $b$ , Somer's  $D$ , and the Pearson and Spearman correlation coefficients. In addition, you can obtain measures of agreement such as the kappa coefficient and the weighted kappa coefficient. McNemar's test is produced for the  $2 \times 2$  table.

The examples in this chapter demonstrate how you can use the Analyst Application to analyze tables, including assessing the presence of association in a table and sets of tables and assessing observer agreement.

---

## Association in a $2 \times 2$ Table

The most basic table is a  $2 \times 2$  table. Usually, the columns represent some sort of outcome, often yes or no, and the rows represent levels of a factor that may influence the outcome. Suppose, for example, that researchers were investigating the properties of a new "ouchless" Band-Aid for children. Interest lies in whether those children trying the test Band-Aid recorded fewer complaints on removal than those children using a regular Band-Aid. You can address this question by forming the two-way table of Band-Aid type and complaint status and then assessing the association between the rows and columns of that table.

### Open the Bandid Data Set

These data are provided as the Bandid data set in the Analyst Sample Library. To open the Bandid data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Bandid.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Bandid from the list of members.
7. Click **OK** to bring the Bandid data set into the data table.

Bandid (Browse)			
	type	outcome	count
1	regular	complain	14
2	regular	no	16
3	test	complain	10
4	test	no	30

**Figure 9.2.** Data Set Bandid in the Data Table

Figure 9.2 displays the data table containing these data. Note that the data are in frequency form, with the variable `count` containing the frequencies of the profile contained in each row of the table. The variable `type` is the type of Band-Aid tested and the variable `outcome` is the status of complaints.

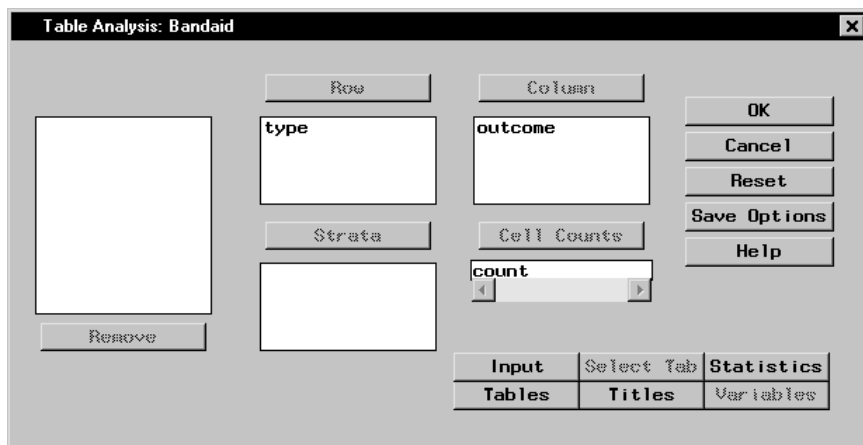
### Specify the Table

To construct the appropriate two-way table and request tests of association, follow these steps:

1. Select **Statistics** → **Table Analysis** . . .
2. Select `type` from the candidate list as the **Row** variable.

3. Select **outcome** from the candidate list as the **Column** variable.
4. Select **count** from the candidate list as the **Cell Counts** variable.

Figure 9.3 displays the resulting dialog.



**Figure 9.3.** Table Analysis Task for Band-Aid Study

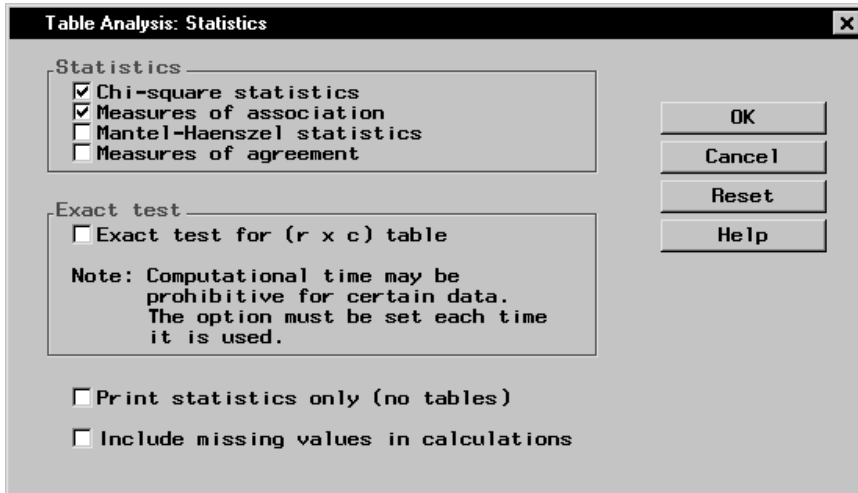
### ***Request Tests and Measures of Association***

By selecting the rows and columns of the table, you have requested the construction of a  $2 \times 2$  table. To request chi-square tests of association and the odds ratio, which is a measure of association, follow these steps:

1. Click on the **Statistics** button.
2. Select **Chi-square statistics**.
3. Select **Measures of association**.
4. Click **OK**.

Figure 9.4 displays the Statistics dialog.



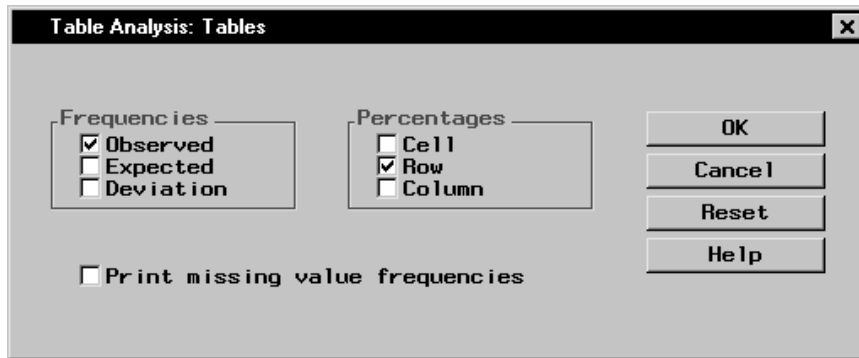


**Figure 9.4.** Statistics Dialog

Finally, in order to customize the form of the displayed table, follow these steps:

1. Click on the **Tables** button.
2. Select **Observed** under **Frequencies**.
3. Select **Row** under **Percentages**.
4. Click **OK**.

Figure 9.5 displays the resulting Tables dialog.



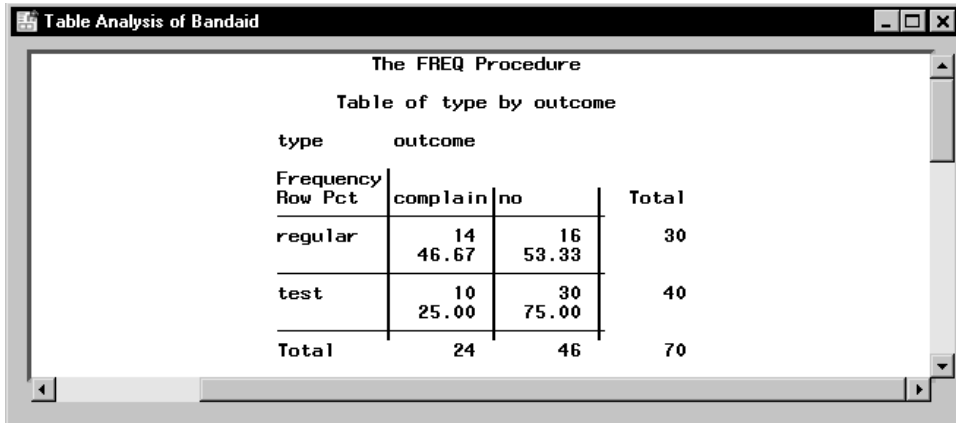
**Figure 9.5.** Tables Dialog

This requests that only the raw frequencies and the row percentages be listed in the printed table cell.

Click **OK** in the Table Analysis dialog to perform the analysis.

### ***Review the Results***

The frequency table is displayed in [Figure 9.6](#). Note that 46 percent of those children getting regular Band-Aids had complaints about irritation when their Band-Aid was removed, compared to 25 percent of those children receiving the test Band-Aid.



The screenshot shows a window titled "Table Analysis of Bandaid" with a sub-header "The FREQ Procedure". Below this is the title "Table of type by outcome". The table displays the relationship between "type" (regular, test) and "outcome" (complain, no). It includes columns for "Frequency", "Row Pct", and "Total".

type	outcome		Total
	complain	no	
regular	14 46.67	16 53.33	30
test	10 25.00	30 75.00	40
Total	24	46	70

**Figure 9.6.** Frequency Table for Bandaid Data

Figure 9.7 contains the table of computed chi-square statistics for this table. The Pearson chi-square statistic, labeled “Chi-Square,” has a value of 3.57 and an associated  $p$ -value of 0.0588 with 1 degree of freedom. If you were doing strict hypothesis testing, you would not reject the hypothesis of no association at the  $\alpha = 0.05$  level of significance. However, researchers in this case found enough evidence in this pilot study to continue looking into the new product.

Statistic	DF	Value	Prob
Chi-Square	1	3.5719	0.0588
Likelihood Ratio Chi-Square	1	3.5655	0.0590
Continuity Adj. Chi-Square	1	2.6749	0.1019
Mantel-Haenszel Chi-Square	1	3.5208	0.0606
Phi Coefficient		0.2259	
Contingency Coefficient		0.2203	
Cramer's V		0.2259	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	14
Left-sided Pr <= F	0.9840
Right-sided Pr >= F	0.0512
Table Probability (P)	0.0351
Two-sided Pr <= P	0.0769

**Figure 9.7.** Chi-Square Statistics for Bandaid Data

Several other chi-square statistics also appear in this output, such as the likelihood ratio chi-square and the Mantel-Haenszel chi-square. These statistics are asymptotically equivalent.

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	2.6250	0.9530	7.2307
Cohort (Col1 Risk)	1.8667	0.9656	3.6085
Cohort (Col2 Risk)	0.7111	0.4865	1.0394

Sample Size = 70

**Figure 9.8.** Odds Ratio for Bandaid Data

Figure 9.8 contains the table of relative risk estimates including the odds ratio, which is labeled “Case-Control.” The odds ratio is the ratio of the odds of having an outcome for one group versus another. When the odds ratio has the value 1, you have equal odds of having the outcome. When the odds ratio is greater than 1, one group has greater odds of an outcome than the other.

The odds ratio has a value of 2.62, which means that the odds of a complaint are 2.62 times higher for those children using the regular Band-Aid than for those using the test Band-Aid.

---

## Exact Test

You may have noticed that the preceding statistical output also included a test called Fisher’s Exact test. When the sample size for the test of association of a table does not meet the usual guidelines (generally 20-25 total observations for a  $2 \times 2$  table, with 80 percent of the table cells having counts greater than 5), an exact test may be a useful strategy.

The following data illustrate where an exact test may be appropriate. A marketing research firm took a sample of members at a health club and asked them a series of questions. They were interested in gathering information that could help their clients decide on audiences to target for new magazines. One of the questions was what activity the member considered his or her primary activity at the club. Another question was whether the member was considering making a major diet change. The researchers were interested in what types of sports magazines in which to place ads for a new food and nutrition magazine.

### Open the Gym Data Set

These data are provided as the Gym data set in the Analyst Sample Library. To open the Gym data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select **Gym**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .

5. Select **Sasuser** from the list of **Libraries**.
6. Select **Gym** from the list of members.
7. Click **OK** to bring the **Gym** data set into the data table.

Figure 9.9 displays the data table containing these data. Note that the data are in frequency form, with the variable **count** containing the frequencies of the profile contained in each row of the table. The variable **activity** contains the type of activity, which can be aerobics, yoga, weightlifting, team sports such as volleyball and basketball leagues, and cross-training. The variable **DietChange** indicates whether the member was contemplating a change in diet.

Gym (Browse)			
	activity	DietChange	count
1	aerobics	yes	13
2	aerobics	no	8
3	yoga	yes	3
4	yoga	no	2
5	weights	yes	3
6	weights	no	19
7	team	yes	12
8	team	no	16
9	cross	yes	11
10	cross	no	13

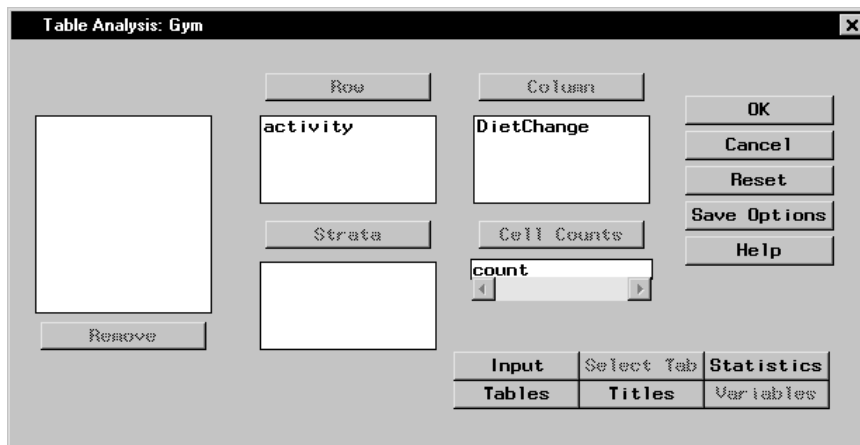
Figure 9.9. Data Set Gym in the Data Table

### Specify the Table

To construct the appropriate two-way table and request tests of association, follow these steps:

1. Select **Statistics** → **Table Analysis** . . .
2. Select **activity** from the candidate list as the **Row** variable.
3. Select **DietChange** from the candidate list as the **Column** variable.
4. Select **count** from the candidate list as the **Cell Counts** variable.

Figure 9.10 displays the resulting dialog.



**Figure 9.10.** Table Analysis Task for Health Club Study

### ***Request Tests and Measures of Association***

By selecting the rows and columns of the table, you have requested the construction of a  $5 \times 2$  table. To request chi-square tests of association, follow these steps:

1. Click on the **Statistics** button.
2. Select **Chi-square statistics**.
3. Click **OK**.

Note that the Tables dialog specifications (see [Figure 9.5](#)) made in the previous analysis remain in effect. Therefore, both frequencies and row percentages are produced for this analysis.

Click **OK** in the Table Analysis dialog to perform the analysis.

**Review the Results**

The frequency table is displayed in [Figure 9.11](#). Note that 62 percent of those members participating in aerobics were considering a diet change and so were 60 percent of yoga practitioners. Eighty-six percent of those members lifting weights were not considering a diet change. Of those members playing a team sport or who considered themselves cross-trainers, the majority of members were not considering a diet change, but not by a wide margin.

The screenshot shows a window titled "Table Analysis of Gym" containing a table titled "The FREQ Procedure" and "Table of activity by DietChange". The table has columns for "activity", "DietChange" (subdivided into "no" and "yes"), and "Total". The rows represent different activities: aerobics, cross, team, weights, and yoga. Each cell contains a count and a percentage.

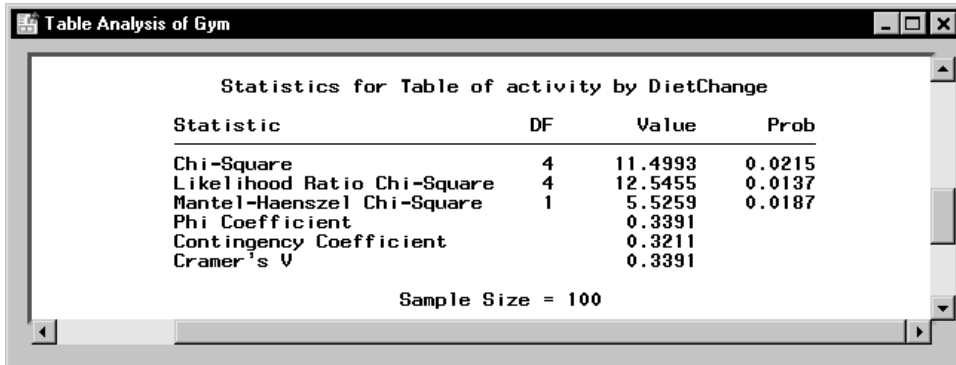
activity	DietChange		Total
	no	yes	
aerobics	8 38.10	13 61.90	21
cross	13 54.17	11 45.83	24
team	16 57.14	12 42.86	28
weights	19 86.36	3 13.64	22
yoga	2 40.00	3 60.00	5
Total	58	42	100

**Figure 9.11.** Frequency Table for Gym Data

[Figure 9.12](#) contains the table of chi-square statistics computed for this table. The Pearson chi-square statistic has a value of 11.4993 and an associated  $p$ -value of 0.0215 with 4 degrees of freedom. If you were doing strict hypothesis testing, you would reject the hypothesis of no association at the  $\alpha = 0.05$  level of significance. However, if you look at [Figure 9.11](#), you see that three table cells have a count of less than 5, which violates one of the sample



size guidelines for the asymptotic tests. Thus, you may want to compute the exact test for these data.



Statistics for Table of activity by DietChange				
Statistic	DF	Value	Prob	
Chi-Square	4	11.4993	0.0215	
Likelihood Ratio Chi-Square	4	12.5455	0.0137	
Mantel-Haenszel Chi-Square	1	5.5259	0.0187	
Phi Coefficient		0.3391		
Contingency Coefficient		0.3211		
Cramer's V		0.3391		

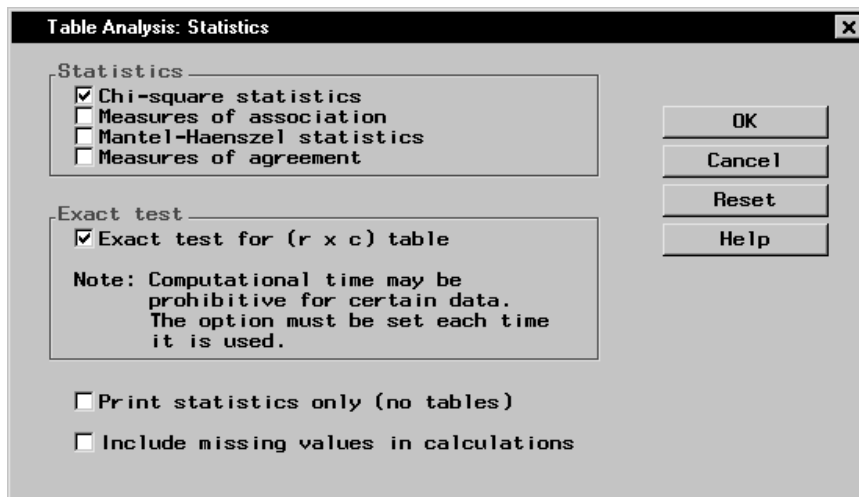
Sample Size = 100

**Figure 9.12.** Chi-square Statistics for Gym Data

### **Request the Exact Test**

To request the exact test, simply return to the Table Analysis task and open the Statistics dialog. All of the settings you have previously selected for the table analysis are still in place. You need only request the additional exact test.

1. Select **Statistics** → **Table Analysis . . .**
2. Click on the **Statistics** button.
3. Select **Exact test for ( $r \times c$ ) table**.
4. Click **OK**.
5. Click **OK** in the main Table Analysis dialog to perform the analysis.



**Figure 9.13.** Statistics Dialog

Figure 9.13 displays the resulting dialog. Notice the warning that exact test computations may take an excessive amount of time. This would not be the case with very small cell counts, but it is an issue for other tables.

### ***Review the Results***

Figure 9.14 contains the results of this analysis, including the exact test results.

Table Analysis of Gym

Statistics for Table of activity by DietChange

Statistic	DF	Value	Prob
Chi-Square	4	11.4993	0.0215
Likelihood Ratio Chi-Square	4	12.5455	0.0137
Mantel-Haenszel Chi-Square	1	5.5259	0.0187
Phi Coefficient		0.3391	
Contingency Coefficient		0.3211	
Cramer's V		0.3391	

Fisher's Exact Test

Table Probability (P)	8.421E-06
Pr <= P	0.0139

Sample Size = 100

**Figure 9.14.** Exact Test Results

The exact test computes a  $p$ -value of 0.0139; thus, this test also results in the rejection of the hypothesis of no association in this table. There is some kind of association between the rows of the table and the columns of the table; type of primary activity made a difference in whether members were considering diet changes. Not only does degree of association seem to vary, but so does the direction. The market research company may end up suggesting that sports and fitness magazines be targeted in different ways for the new food and diet magazine ad campaign.

## Association in Sets of Tables

After the pilot study on the new ouchless Band-Aids, the investigators decided to continue their research by conducting a clinical trial in which children at five clinics were tested with the test and regular Band-Aids. Instead of a single table, the clinical trial produces five tables. In order to assess whether the test Band-Aids produced fewer complaints than the regular Band-Aids, you need to assess the association in sets of tables instead of the association in a single table.

Extended Mantel-Haenszel statistics, also known as Cochran-Mantel-Haenszel statistics, provide a way of assessing association between two variables that determine a table while controlling for, or adjusting for, the variables that determine the sets of tables. These variables are also known as stratification variables. In this instance, the statistics can provide a way to assess the association between Band-Aid type and complaint status while controlling for clinic.

In the first section, the odds ratio was presented as a measure of association. You can also compute an overall odds ratio for a set of tables that has been adjusted for the stratification variables.

The `Studybandaid` data set contains the information collected in this clinical trial and includes data that constitute tables for each of the five clinics.

### **Open the Studybandaid Data Set**

These data are provided as the `Studybandaid` data set in the Analyst Sample Library. To open the `Studybandaid` data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select `Studybandaid`.
3. Click **OK** to create the sample data set in your `Sasuser` directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select `Sasuser` from the list of **Libraries**.
6. Select `Studybandaid` from the list of members.
7. Click **OK** to bring the `Studybandaid` data set into the data table.

[Figure 9.15](#) displays the data table containing these data. Note that the data are in frequency form, with the variable `COUNT` containing the frequencies of the profile contained in each row of the table. The column corresponding to the variable `clinic` contains the values for the five clinics.

Studybandaid (Browse)				
	clinic	type	outcome	count
1	A	regular	complain	14
2	A	regular	no	17
3	A	test	complain	11
4	A	test	no	31
5	B	regular	complain	22
6	B	regular	no	21
7	B	test	complain	10
8	B	test	no	40
9	C	regular	complain	22
10	C	regular	no	28
11	C	test	complain	15
12	C	test	no	30
13	D	regular	complain	15
14	D	regular	no	18
15	D	test	complain	8
16	D	test	no	29
17	E	regular	complain	20
18	E	regular	no	30
19	E	test	complain	15
20	E	test	no	29

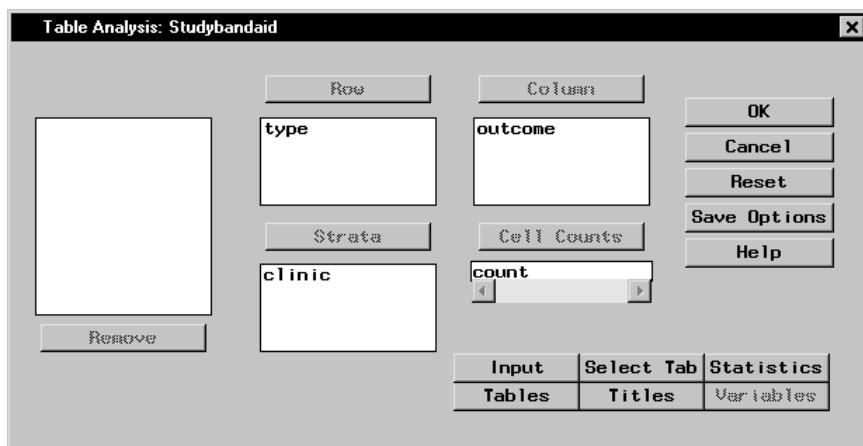
**Figure 9.15.** Data Set Studybandaid in the Data Table

### **Specify the Tables**

To request individual table tests of association as well as the CMH tests for the association of type of Band-Aid with complaint outcome, first specify the tables under study.

1. Select **Statistics** → **Table Analysis . . .**
2. Select **type** from the candidate list as the **Row** variable.
3. Select **outcome** from the candidate list as the **Column** variable.
4. Select **clinic** from the candidate list as the **Strata** variable.
5. Select **count** from the candidate list as the **Cell Counts** variable.

Figure 9.16 displays the resulting dialog.



**Figure 9.16.** Table Analysis Task for Band-Aid Study

### ***Request Tests and Measures of Association***

Use the Statistics dialog to specify the tests.

1. Click on the **Statistics** button.
2. Select **Chi-square statistics**.
3. Select **Mantel-Haenszel Statistics**.
4. Click **OK**.

Note that the Tables dialog specifications (see Figure 9.5) made previously remain in effect. Therefore, both frequencies and row percentages are produced for this analysis.

Click **OK** in the Table Analysis dialog to perform the analysis.

### Review the Results

The results produced include individual tables, individual table statistics, and the summary chi-square statistics.

Table 1 of type by outcome  
Controlling for clinic=A

type	outcome		Total
	complain	no	
regular	14 45.16	17 54.84	31
test	11 26.19	31 73.81	42
<b>Total</b>	<b>25</b>	<b>48</b>	<b>73</b>

**Figure 9.17.** Frequency Table for Clinic A

Figure 9.17 contains the frequency table for clinic A.

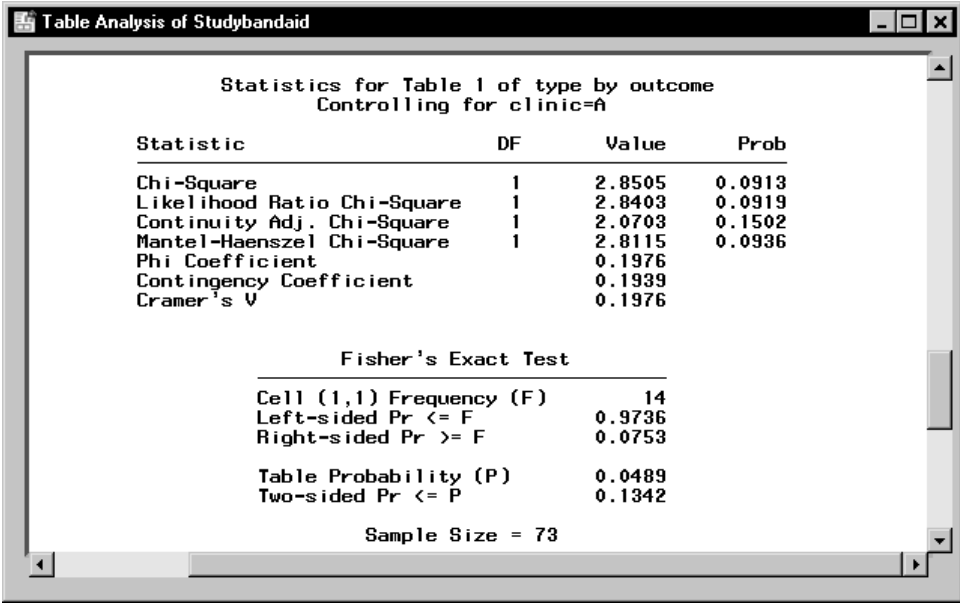


Table Analysis of Studybandaid

Statistics for Table 1 of type by outcome  
Controlling for clinic=A

Statistic	DF	Value	Prob
Chi-Square	1	2.8505	0.0913
Likelihood Ratio Chi-Square	1	2.8403	0.0919
Continuity Adj. Chi-Square	1	2.0703	0.1502
Mantel-Haenszel Chi-Square	1	2.8115	0.0936
Phi Coefficient		0.1976	
Contingency Coefficient		0.1939	
Cramer's V		0.1976	

Fisher's Exact Test

Cell (1,1) Frequency (F)	14
Left-sided Pr <= F	0.9736
Right-sided Pr >= F	0.0753
Table Probability (P)	0.0489
Two-sided Pr <= P	0.1342

Sample Size = 73

**Figure 9.18.** Table Statistics for Clinic A

Figure 9.18 contains the table statistics for clinic A. The Pearson chi-square statistic has the value 2.8505 and a  $p$ -value of 0.091 with 1 degree of freedom.



Table 2 of type by outcome  
Controlling for clinic=B

type	outcome		Total
Frequency	complain	no	
Row Pct			
regular	22 51.16	21 48.84	43
test	10 20.00	40 80.00	50
Total	32	61	93

**Figure 9.19.** Frequency Table for Clinic B

Figure 9.19 contains the frequency table for clinic B.

Table Analysis of Studybandaid

Statistics for Table 2 of type by outcome  
Controlling for clinic=B

Statistic	DF	Value	Prob
Chi-Square	1	9.9475	0.0016
Likelihood Ratio Chi-Square	1	10.1022	0.0015
Continuity Adj. Chi-Square	1	8.6146	0.0033
Mantel-Haenszel Chi-Square	1	9.8405	0.0017
Phi Coefficient		0.3271	
Contingency Coefficient		0.3108	
Cramer's V		0.3271	

Fisher's Exact Test

Cell (1,1) Frequency (F)	22
Left-sided Pr <= F	0.9997
Right-sided Pr >= F	0.0016
Table Probability (P)	0.0012
Two-sided Pr <= P	0.0022

Sample Size = 93

**Figure 9.20.** Table Statistics for Clinic B

Figure 9.20 contains the associated table statistics. The Pearson chi-square statistic has a value of 9.9475 and a corresponding  $p$ -value of 0.0016.

The other individual tables, not printed here, show varying degrees of evidence of association. Clinic C and clinic E appear to have no evidence of association, while clinic D does appear to show evidence of association.

Summary Statistics for type by outcome  
Controlling for clinic

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	14.2206	0.0002
2	Row Mean Scores Differ	1	14.2206	0.0002
3	General Association	1	14.2206	0.0002

**Figure 9.21.** CMH Summary Table

Figure 9.21 displays the results of the CMH analysis. Three versions of the CMH statistic are printed; all have the value 14.2206 and a  $p$ -value of 0.0002 with 1 degree of freedom. Your choice of statistic depends on the scale of variables that determine the rows and columns. The General Association statistic always applies. If the columns can be considered ordered, or ordinal, then the Row Mean Score statistic is appropriate as well and is directed at location shifts. If both the columns and rows are ordered, then the Correlation statistic is also appropriate and is directed at linear association. The degrees of freedom of these statistics vary. For more details, refer to Stokes, Davis, and Koch (1995). Note that the sample size requirement for the CMH statistics is that the total (tables combined) sample size be adequate.

In the case of the  $2 \times 2$  table, all of these statistics are equivalent. Here, you can conclude that type of Band-Aid is significantly associated with complaint status, controlling for clinic. Figure 9.22 displays the overall relative risk and odds ratios and their confidence bounds.

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	2.1597	1.4420	3.2348
	Logit	2.1561	1.4331	3.2439
Cohort (Co11 Risk)	Mantel-Haenszel	1.6446	1.2637	2.1402
	Logit	1.6112	1.2355	2.1013
Cohort (Co12 Risk)	Mantel-Haenszel	0.7563	0.6510	0.8787
	Logit	0.7606	0.6545	0.8838

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square	4.4750
DF	4
Pr > ChiSq	0.3455

Total Sample Size = 425

**Figure 9.22.** Odds Ratio

The odds ratio for this study has the value 2.1597 with a confidence bound of (1.4420, 3.2348). This means that those children with the regular Band-Aid are twice as likely to have complaints as those with the test Band-Aid or, conversely, that those children with the test Band-Aid are half as likely to have complaints as those children with the regular Band-Aid. Since the 95 percent confidence bounds don't include the value 1, this odds ratio is considered to be significantly different from 1.

Note that another test called the Breslow-Day test for Homogeneity of Odds Ratio is also printed. Since the test has a  $p$ -value of 0.3455, you would conclude that the hypothesis is not rejected. The sample size requirement for this test is that each individual table has to have sufficient sample size unlike the sample size requirement for the CMH statistics. In this case, since all tables have totals greater than 25, this condition is met.

---

## Observer Agreement

Often, the data represented by a contingency table represents information collected in a study on observer agreement. There may be interest in gathering information on observer error, and such a study may be done as part of testing new processes, training, or tools. Sometimes different observers are studied, and sometimes the same observer is studied at different times or under different conditions.

The members of a northeastern music association were revising their system of conducting local and state-wide high school piano competitions. Instead of using local musicians as judges, they wanted to see if they could proceed more fairly by using one of two trained judges in conjunction with local judges, with whom they needed to come to consensus. In order to see how closely the trained judges match, they did an observer agreement study using some college music students after a training session. Twenty students played one of their current pieces, and both judges rated the performance as good, skilled, or superior.

In order to analyze such data, you form the table with the ratings of one rater forming the rows of the table and the ratings of the other rater forming the columns of the table. The cells of the table are the number of students who fell into the profiles composed of the combination of both ratings. Since there are 3 outcomes, there are 9 possible combinations as represented by the cells of a two-way table. Statistics called measures of agreement are then used to assess the degree of agreement.

### *Open the Piano Data Set*

The Piano data set contains the variables Rater1 and Rater2 as well as a frequency variable count. These data are provided as the Piano data set in the Analyst Sample Library. To open the Piano data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Piano.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .

5. Select **Sasuser** from the list of **Libraries**.
6. Select **Piano** from the list of members.
7. Click **OK** to bring the Piano data set into the data table.

Figure 9.23 displays the data table containing these data. Note that the data are in frequency form, with the variable **count** containing the frequencies of the profile contained in each row of the table. The variable **Rater1** contains the first rater's evaluations and the variable **Rater2** contains the second rater's evaluations.

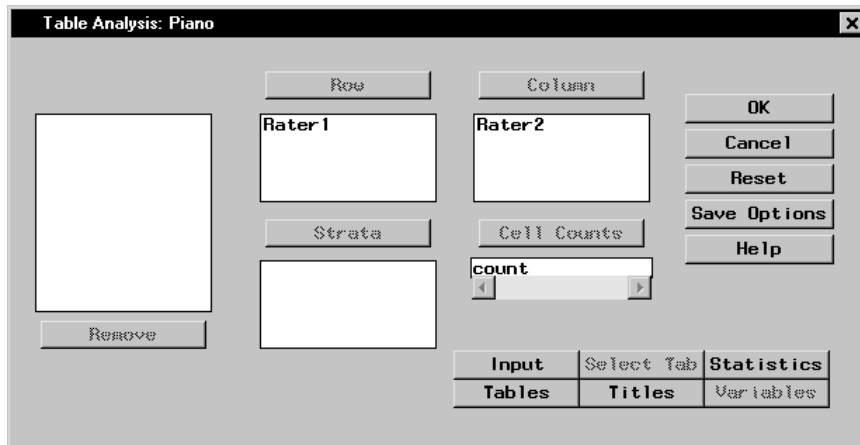
Piano (Browse)			
	Rater1	Rater2	count
1	good	good	5
2	good	skilled	1
3	good	superior	0
4	skilled	good	2
5	skilled	skilled	5
6	skilled	superior	2
7	superior	good	1
8	superior	skilled	1
9	superior	superior	3

Figure 9.23. Data Set Piano in the Data Table

### Specify the Table

To construct the appropriate two-way table, follow these steps:

1. Select **Statistics** → **Table Analysis** . . .
2. Select **Rater1** from the candidate list as the **Row** variable.
3. Select **Rater2** from the candidate list as the **Column** variable.
4. Select **count** from the candidate list as the **Cell Counts** variable.



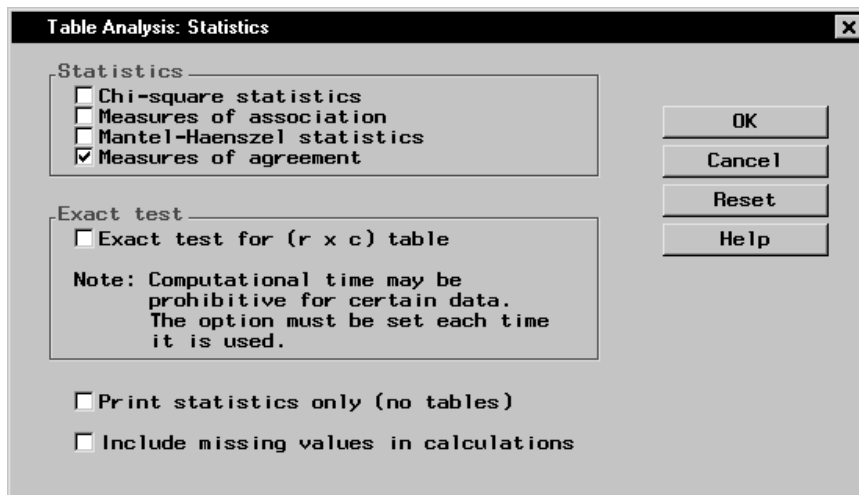
**Figure 9.24.** Table Analysis Task for Music Study

Figure 9.24 displays the resulting dialog.

### ***Request Measures of Agreement***

To request measures of agreement, follow these steps:

1. Click on the **Statistics** button.
2. Select **Measures of agreement**.
3. Click **OK**.



**Figure 9.25.** Statistics Dialog

Figure 9.25 displays the resulting Statistics dialog. Note that the chi-square tests of association and the measures of association are not appropriate for this type of table.

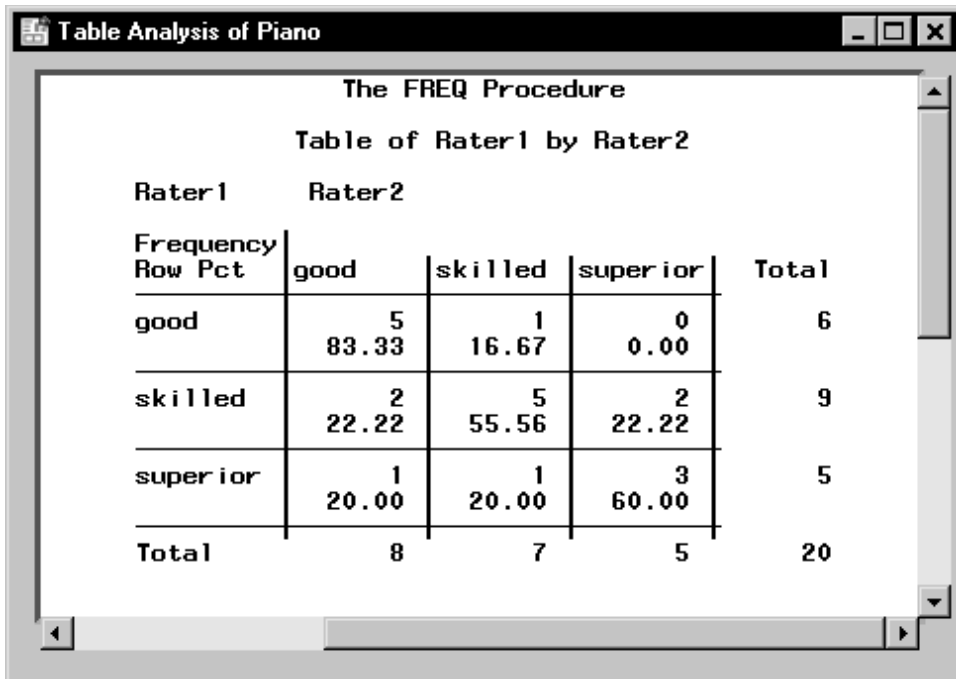
Note that the Tables dialog specifications (see Figure 9.5) made previously remain in effect. Therefore, both frequencies and row percentages are produced for this analysis.

Click **OK** in the Table Analysis dialog to perform the analysis.

### **Review the Results**

The frequency table is displayed in Figure 9.26. Note that most of the frequencies occur on the diagonals, which is what you would expect if there is any degree of agreement. However, there are several off-diagonal elements that represent nonagreement. In particular, there is one case of a student rated ‘good’ by Rater2 and ‘superior’ by Rater1. This might be unexpected.





The screenshot shows a window titled "Table Analysis of Piano" with a standard Windows-style title bar. Inside the window, the text "The FREQ Procedure" is centered at the top. Below it, the title "Table of Rater1 by Rater2" is displayed. The table itself is a 4x4 grid with the following data:

Rater1	Rater2			
Frequency Row Pct	good	skilled	superior	Total
good	5 83.33	1 16.67	0 0.00	6
skilled	2 22.22	5 55.56	2 22.22	9
superior	1 20.00	1 20.00	3 60.00	5
Total	8	7	5	20

Figure 9.26. Piano Agreement Frequency Table

Table Analysis of Piano

Statistics for Table of Rater1 by Rater2

Test of Symmetry

Statistic (S)	1.6667
DF	3
Pr > S	0.6444

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits
Simple Kappa	0.4697	0.1597	0.1566 0.7828
Weighted Kappa	0.5210	0.1563	0.2147 0.8272

Sample Size = 20

**Figure 9.27.** Measures of Agreement

Figure 9.27 contains the results for the measures of agreement. The simple kappa coefficient has a value of 0.4697, with a 95 percent confidence bounds of (0.1566, 0.7828). This suggests modest agreement of ratings. Note that the Bowker's test of symmetry is also printed; this is a test that the probabilities represented by a square table satisfy symmetry.

When you have a  $2 \times 2$  table, the measure of agreement produced is McNemar's test.

---

## References

- SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.
- Stokes, Maura E., Davis, Charles S., and Koch, Gary G. (1995), *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc.

# Chapter 10

## Analysis of Variance

### Chapter Contents

---

<b>Introduction</b> . . . . .	269
<b>One-Way Analysis of Variance</b> . . . . .	273
<b>Nonparametric One-Way Analysis of Variance</b> . . . . .	279
<b>Factorial Analysis of Variance</b> . . . . .	284
<b>Linear Models</b> . . . . .	290
<b>References</b> . . . . .	298



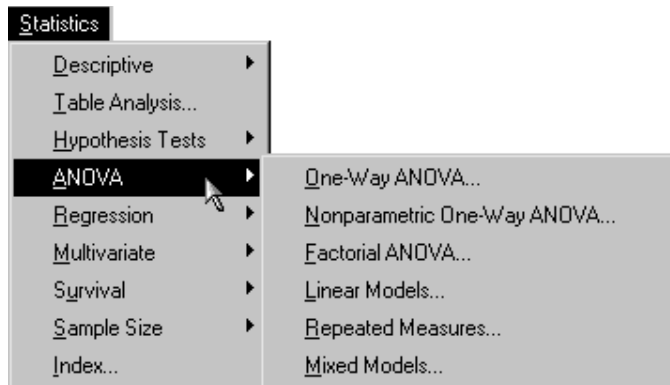
# Chapter 10

## Analysis of Variance

---

### Introduction

Analysis of variance is a technique for exploring the variation of a continuous response variable (dependent variable). The response variable is measured at different levels of one or more classification variables (independent variables). The variation in the response due to the classification variables is computed, and you can test this variation against the residual error to determine the significance of the classification effects.



**Figure 10.1.** Analysis of Variance Menu

The Analyst Application provides several types of analyses of variance (ANOVA). The One-Way ANOVA task compares the means of the response variable over the groups defined by a single classification variable. See the section “[One-Way Analysis of Variance](#)” beginning on page 273 for more information.

The Nonparametric One-Way ANOVA task performs tests for location and scale differences over the groups defined by a single classification variable.

Eight nonparametric tests are offered. See the section “[Nonparametric One-Way Analysis of Variance](#)” beginning on page 279 for more information.

The Factorial ANOVA task compares the means of the response variable over the groups defined by one or more classification variables. This type of analysis is useful when you have multiple ways of classifying the response values. See the “[Factorial Analysis of Variance](#)” section beginning on page 284 for more information.

The Linear Models task enables you to compare means and explain variation when you have a model that includes classification variables, quantitative variables, or both (such as in an analysis of covariance). See the “[Linear Models](#)” section beginning on page 290 for more information.

You can use the Repeated Measures task when you have multiple measurements of the response variable for the same experimental unit over different times or conditions or when the response values are assumed to be correlated within certain groups. For detailed information, see [Chapter 16, “Repeated Measures.”](#)

The Mixed Models task enables you to fit basic mixed models. A mixed model is a linear model that contains both fixed effects and random effects. For detailed information, see [Chapter 15, “Mixed Models.”](#)

The examples in this chapter demonstrate how you can use the Analyst Application to perform one-way and factorial ANOVA as well as to fit the linear model.

### ***The Air Quality Data Set***

The data set used in the following examples contains measurements on air quality recorded in an industrial valley. The measurements are taken hourly for a period of one week.

The first variable in the data set `Air` is a SAS datetime variable (`datetime`) that contains the date and the time of day on which the observation was taken. The data set contains two additional time-related variables related to `datetime` that record the day of the week (`day`) and the hour of the day (`hour`).

The variables measuring air quality are CO (carbon monoxide), O3 (ozone), SO4 (sulfate), NO (nitrous oxide), and dust (particulates). The final variable provided is wind, which gives the wind speed in knots.

### **Open the Air Data Set**

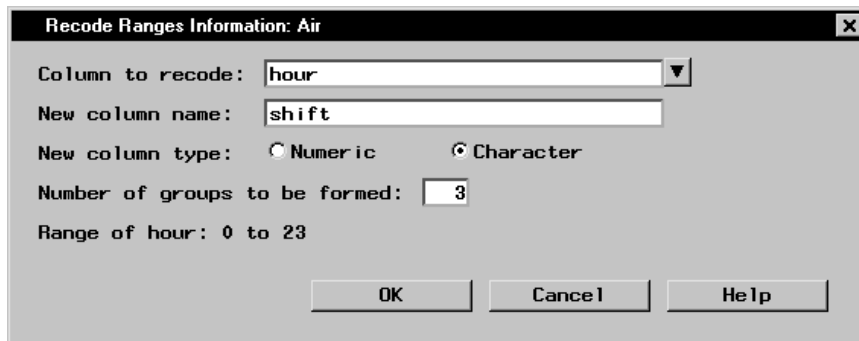
The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select **Air**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Air** from the list of members.
7. Click **OK** to bring the **Air** data set into the data table.

### **Create a New Variable**

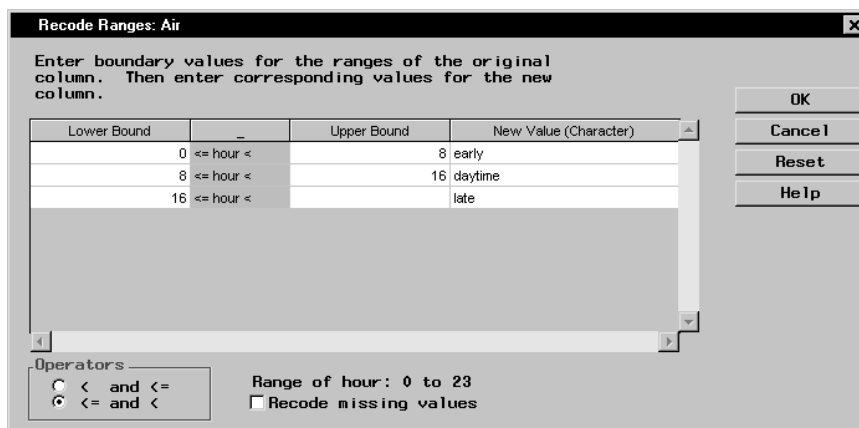
To perform the analyses in the following examples, you need to create a new variable to represent the factory workshift periods. The new character variable, **shift**, recodes the variable **hour** into three factory workshift periods. For information on recoding ranges and computing variables, see the section “[Recoding Ranges](#)” on page 44 in [Chapter 2](#).

[Figure 10.2](#) displays the Recoding Ranges Information dialog. Enter the information to create the new variable as shown in [Figure 10.2](#).



**Figure 10.2.** Recoding Ranges Information: Defining the New Variable

Click **OK** to display the Recoding Ranges dialog (Figure 10.3). To define the values for the new variable, shift, enter the values as shown in Figure 10.3.



**Figure 10.3.** Recoding Ranges: Defining the Values for the New Variable

The values of the new variable Shift are as follows: ‘early’ corresponds to the hours between 0 and 8 (from midnight until 8 a.m.), ‘daytime’ corresponds to the hours between 8 and 16 (from 8 a.m. until 4 p.m.), and ‘late’ corresponds to the hours greater than or equal to 16 (from 4 p.m. to midnight).



---

## One-Way Analysis of Variance

The One-Way ANOVA task enables you to perform an analysis of variance when you have a continuous dependent variable and a single classification variable.

For example, consider the data set on air quality (Air), described in the preceding section. Suppose you want to compare the ozone level corresponding to each of the three factory workshift periods.

### Request the One-Way ANOVA Task

To request the one-way ANOVA task, follow these steps:

1. Select **Statistics** → **ANOVA** → **One-Way ANOVA . . .**
2. Select **o3** as the dependent variable.
3. Select **shift** as the independent variable.

Figure 10.4 defines the one-way ANOVA model.

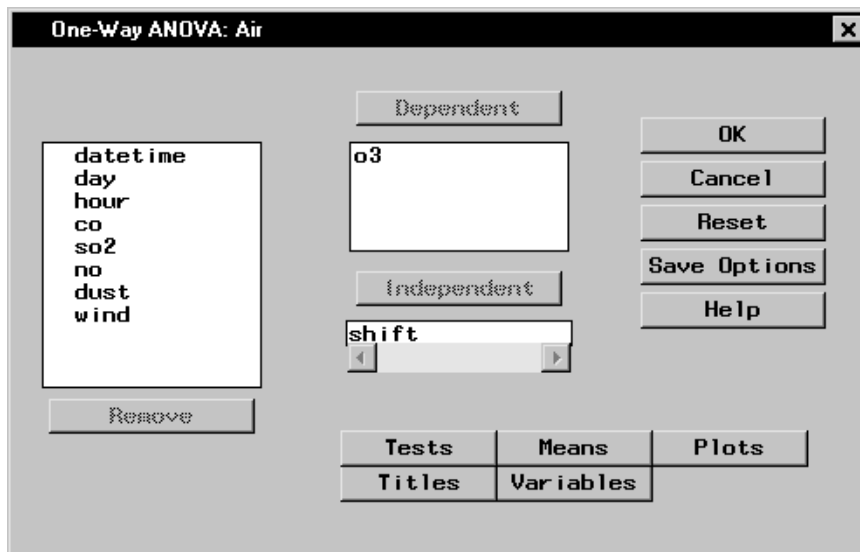


Figure 10.4. One-Way ANOVA Dialog

### **Request a Means Comparison Test**

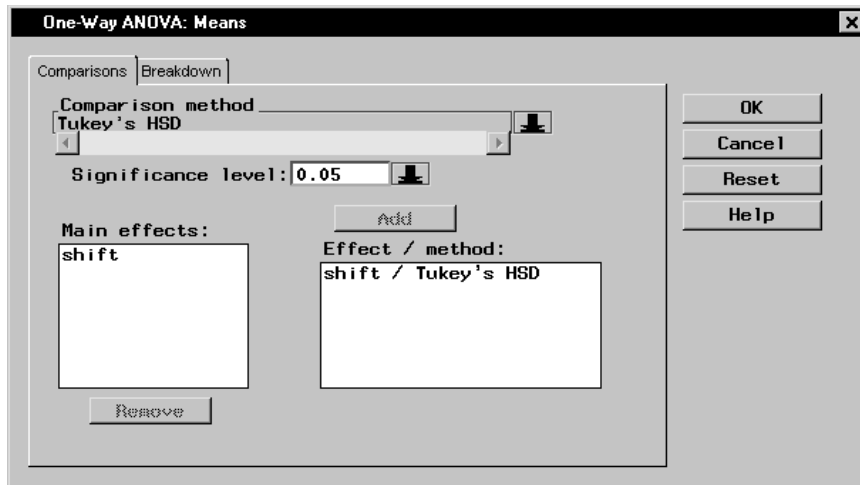
The analysis of variance performed in the One-Way ANOVA task indicates whether the means of the groups are different; it does not indicate which particular means are different. To generate more detailed information about the differences between the means, follow these steps:

1. Click on the **Means** button in the main dialog. The resulting window displays the **Comparisons** tab.
2. Click on the arrow adjacent to the **Comparison method** list.
3. Select **Tukey's HSD**.
4. Highlight the variable **shift** in the **Main Effects:** box.
5. Click on the **Add** button.

You can click on the arrow next to **Significance level:** to select a significance level, or you can type in the desired value.

6. Click **OK**.

**Figure 10.5** specifies Tukey's studentized range (HSD) means comparison test at the 0.05 significance level.



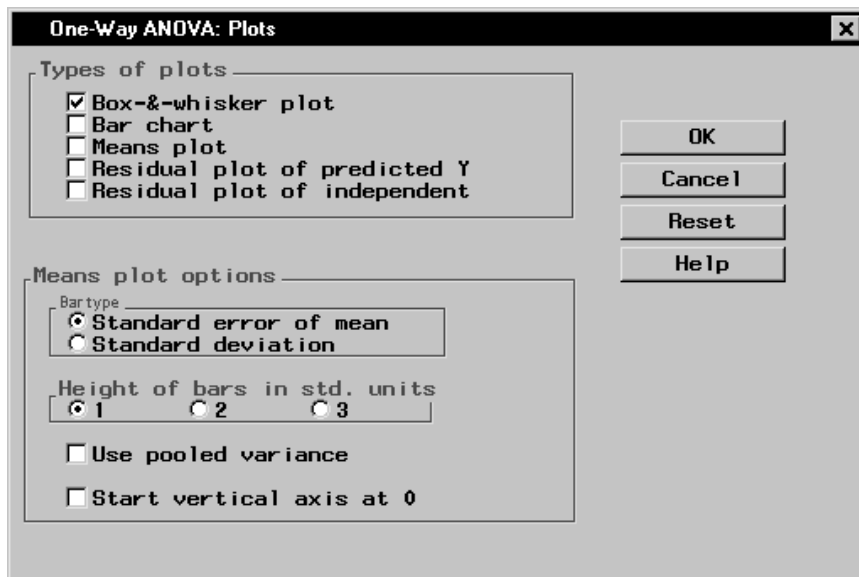
**Figure 10.5.** One-Way ANOVA: Means Dialog

### ***Request a Box-and-Whisker Plot***

To request a box-and-whisker plot in addition to the analysis, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Box-&-whisker plot**.
3. Click **OK**.

Figure 10.6 displays the Plots dialog with the **Box-&-whisker plot** selected.

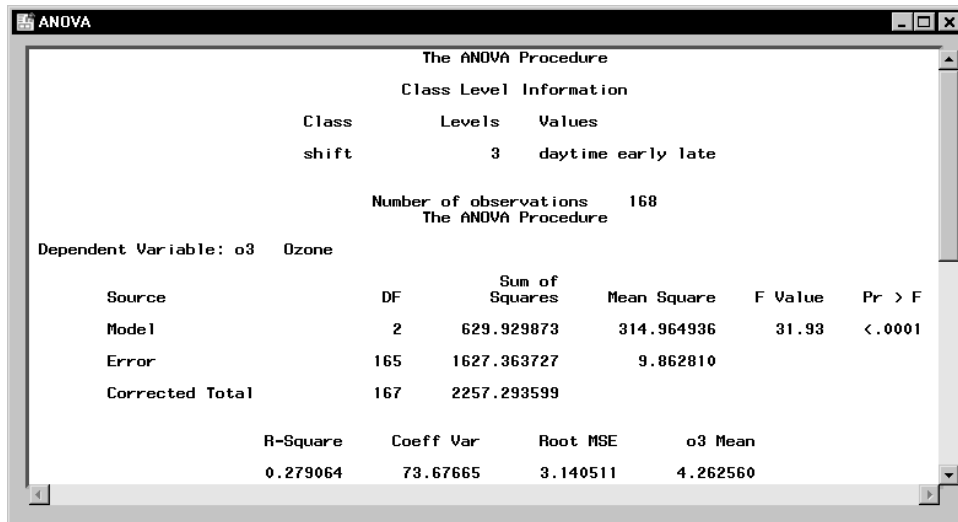


**Figure 10.6.** One-Way ANOVA: Plots Dialog

Click **OK** in the One-Way ANOVA dialog to perform the analysis.

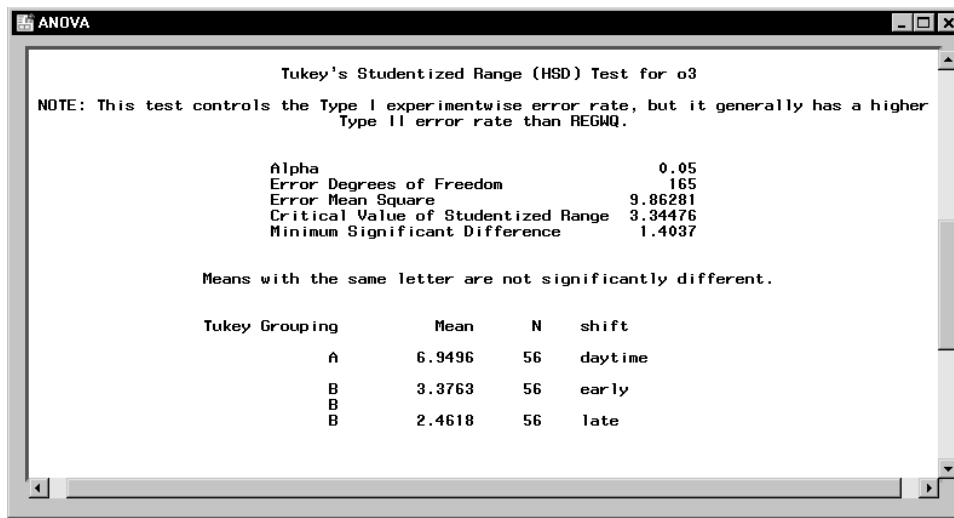
### **Review the Results**

This analysis tests whether the independent variable (shift) is a significant factor in accounting for the variation in ozone levels. [Figure 10.7](#) displays the analysis of variance table, with an  $F$  statistic of 31.93 and an associated  $p$ -value that is less than 0.0001. The small  $p$ -value indicates that the model explains a highly significant proportion of the variation present in the dependent variable.



**Figure 10.7.** One-Way ANOVA: Analysis Results

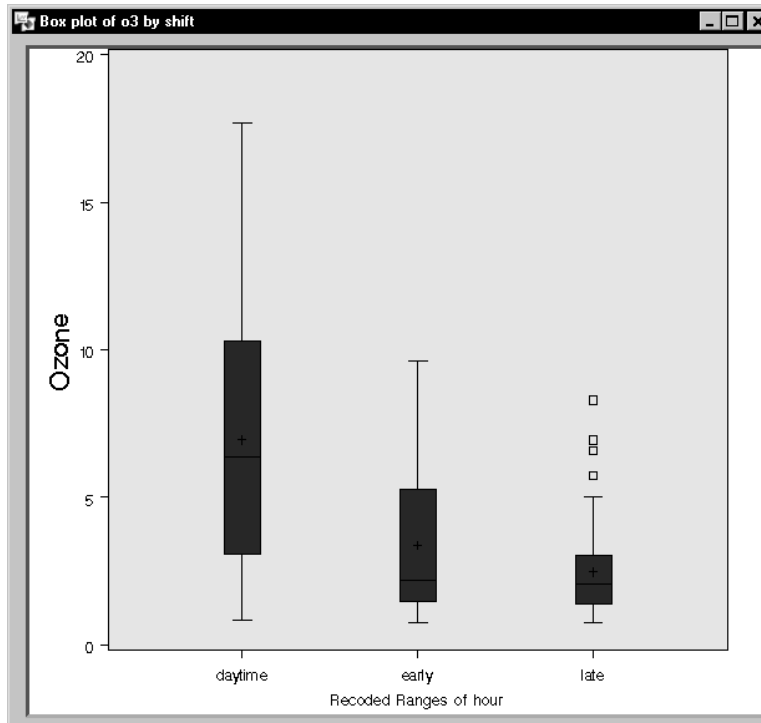
The R-square value, which follows the ANOVA table in [Figure 10.7](#), represents the proportion of variability accounted for by the independent variable. Approximately 28% of the variability in the ozone level can be accounted for by differences between shifts.



**Figure 10.8.** One-Way ANOVA: Multiple Comparisons Results

Information detailing which particular means are different is available in the multiple comparison test, as displayed in [Figure 10.8](#). The means comparison output provides the alpha value, error degrees of freedom, and error mean square.

In the “Tukey Grouping” table, means with the same letter are not significantly different. The analysis shows that the daytime shift is associated with ozone levels that are significantly different from the other two shifts. The early and late shifts cannot be statistically distinguished on the basis of mean ozone level.



**Figure 10.9.** One-Way ANOVA: Box-and-Whisker Plot

The box-and-whisker plot displayed in Figure 10.9 provides a graphical view of the multiple comparison results. The variance among the ozone levels may be unequal: subsequent analyses may include a test for homogeneity of variance or a transformation of the response variable,  $o_3$ .

---

## Nonparametric One-Way Analysis of Variance

In statistical inference, or hypothesis testing, the traditional tests are called parametric tests because they depend on the specification of a probability distribution (such as the normal) except for a set of free parameters. Parametric tests are said to depend on distributional assumptions. Nonparametric tests, on the other hand, do not require distributional assumptions. Even if the data

are distributed normally, nonparametric methods are often almost as powerful as parametric methods.

The Nonparametric One-Way ANOVA task enables you to perform nonparametric tests for location and scale when you have a continuous dependent variable and a single independent classification variable. You can perform a nonparametric one-way ANOVA using Wilcoxon (Kruskal-Wallis), median, Van der Waerden, and Savage scores. In addition, you can test for scale differences across levels of the independent variable using Ansari-Bradley, Siegal-Tukey, Klotz, and Mood scores. The Nonparametric One-Way ANOVA task provides asymptotic and exact  $p$ -values for all tests for location and scale.

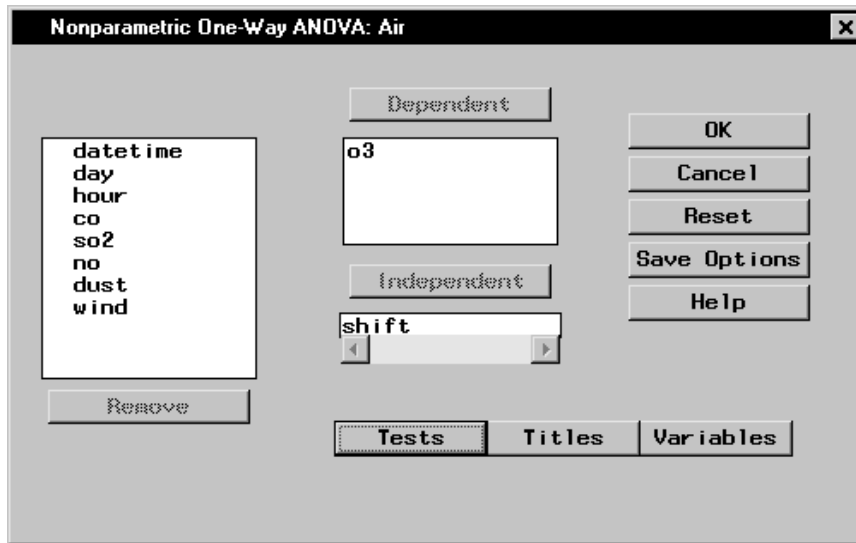
For example, consider the air quality data set (Air), described in the section “[The Air Quality Data Set](#)” on page 270. Suppose that you want to perform a nonparametric one-way ANOVA and also test for scale differences for ozone levels across shift periods.

### **Request the Nonparametric One-Way ANOVA**

To request a nonparametric one-way ANOVA, follow these steps:

1. Select **Statistics** → **ANOVA** → **Nonparametric One-Way ANOVA** . . .
2. Select **o3** as the dependent variable.
3. Select **shift** as the independent variable.





**Figure 10.10.** Nonparametric One-Way ANOVA: Main Dialog

Figure 10.10 defines the nonparametric one-way ANOVA model.

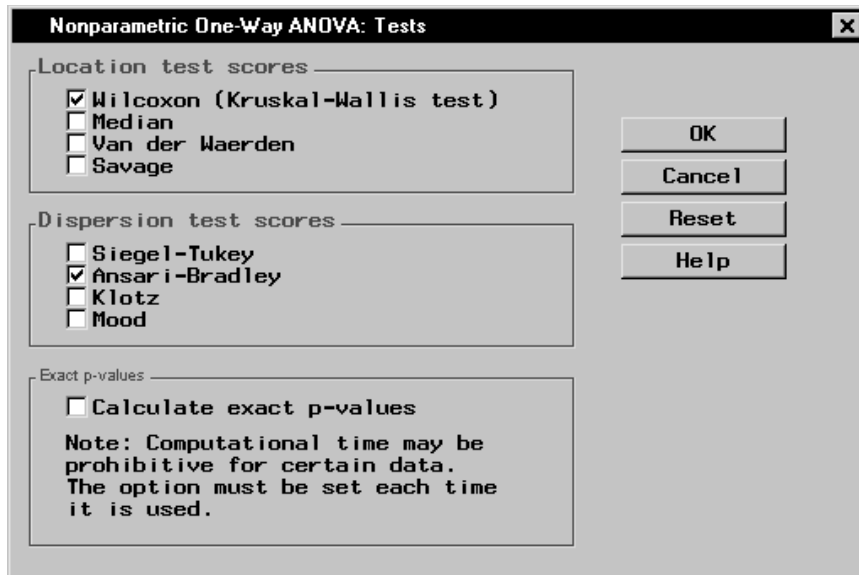
### **Request Nonparametric Tests**

You can use a nonparametric test for location to determine whether the air quality is the same at different times of the day. The Kruskal-Wallis test is a commonly used nonparametric technique for testing location differences and is produced using Wilcoxon scores.

The box-and-whisker plot in Figure 10.9 indicates that ozone levels may be more variable during the daytime shift than during the early shift or at night. You can use the Ansari-Bradley test to test for scale differences across shifts.

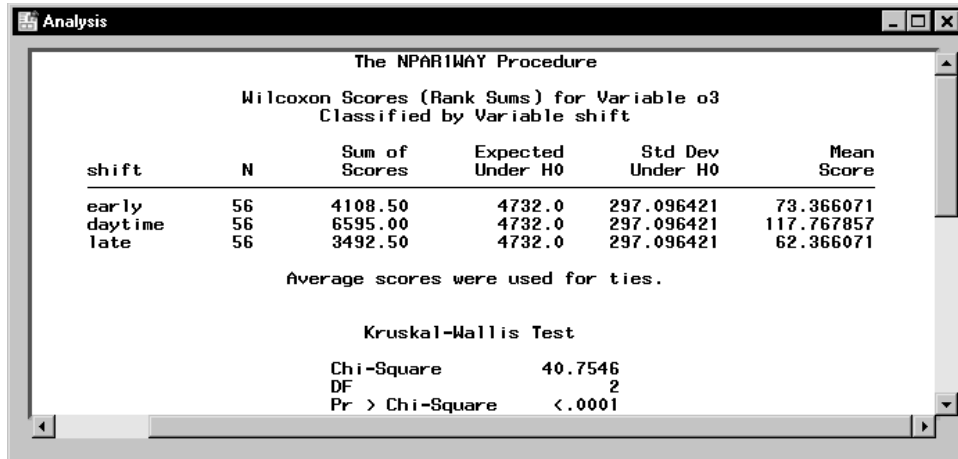
To request the Kruskal-Wallis and Ansari-Bradley tests, follow these steps:

1. Click on the **Tests** button in the main dialog.
2. Select **Wilcoxon (Kruskal-Wallis test)** in the **Location test scores**.
3. Select **Ansari-Bradley** in the **Dispersion test scores** box.



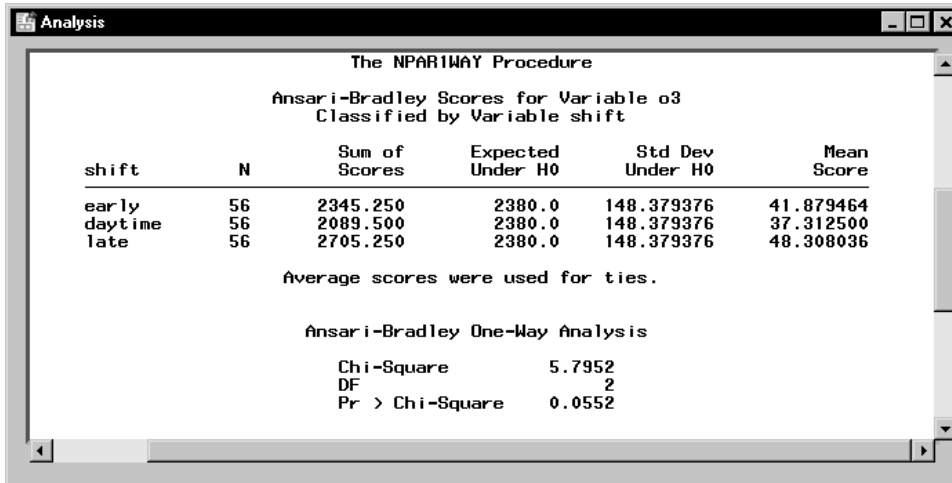
**Figure 10.11.** Nonparametric One-Way ANOVA: Tests Dialog

Figure 10.11 displays the Tests dialog with the **Wilcoxon (Kruskal-Wallis)** and **Ansari-Bradley** tests selected. Click **OK** in the Nonparametric One-Way ANOVA dialog to perform the analysis.



**Figure 10.12.** Nonparametric One-Way ANOVA: Kruskal-Wallis Test Results

Figure 10.12 displays the Wilcoxon scores and Kruskal-Wallis test results. The table labeled “Wilcoxon Scores (Rank Sums) for Variable o3” contains the sum of the rank scores, expected sum, and mean score for each shift. The daytime shift has a mean score of 117.77, which is higher than the mean scores of both the early and late shift. The “Kruskal-Wallis Test” table displays the results of the Kruskal-Wallis test. The test statistic of 40.75 indicates that there is a significant difference in ozone levels across shift times (the  $p$ -value is less than 0.0001).



**Figure 10.13.** Nonparametric One-Way ANOVA: Ansari-Bradley Test Results

Figure 10.13 displays the results of the Ansari-Bradley test. The Ansari-Bradley test chi-square has the value of 5.80 with 2 degrees of freedom, which is not significant at the  $\alpha = 0.05$  level. Since the  $p$ -value is just slightly higher than 0.05, there is moderate evidence of scale differences across shift times.

## Factorial Analysis of Variance

The Factorial ANOVA task enables you to perform an analysis of variance when you have multiple classification variables.

For example, consider the data set on air quality (Air), described in the section “The Air Quality Data Set” on page 270. Suppose you want to compare ozone levels for each day of the week and for each factory workshift. You can define a factorial model that includes the two classification variables, *day* and *shift*.

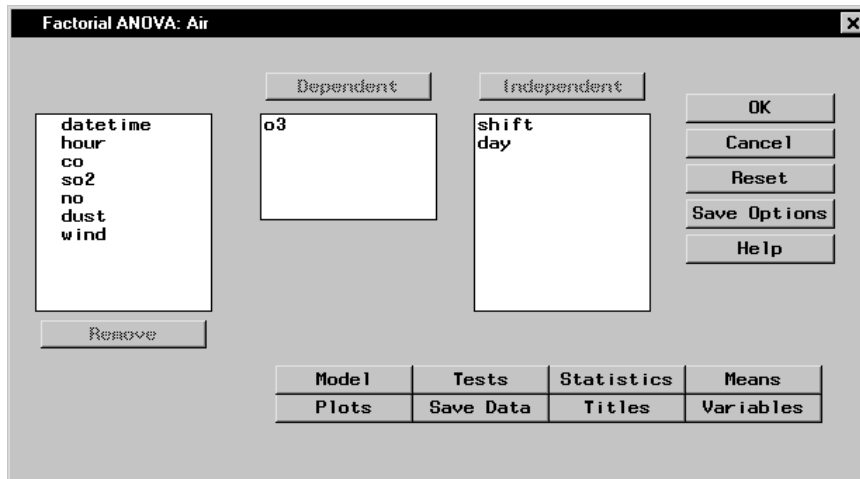
In this example, a factorial model is specified, and a plot of the two-way effects is requested.

### Request the Analysis

To request a factorial analysis of variance, follow these steps:

1. Click on **Statistics** → **ANOVA** → **Factorial ANOVA . . .**
2. Select **o3** as the dependent variable.
3. Select **shift** and **day** as the independent variables.

The resulting Factorial ANOVA dialog is displayed in [Figure 10.14](#).



**Figure 10.14.** Factorial ANOVA Dialog

The default ANOVA model includes only the main effects (that is, the terms representing **shift** and **day**). To include an interaction term, or to specify other options for your analysis, you can use the dialogs available in the Factorial ANOVA task.

### Specify the Model

To specify a factorial model, follow these steps:

1. Click on the **Model** button in the main dialog.
2. Highlight the variables **shift** and **day** in the resulting dialog.
3. Click on the **Factorial** button.
4. Click **OK**.

Figure 10.15 displays the Model dialog with the terms **shift**, **day**, and the interaction term **shift\*day** selected as effects in the model.

Note that you can build specific models with the **Add**, **Cross**, and **Factorial** buttons, or you can select a model by clicking on the **Standard Models** button and making a selection from the drop-down list. From this list, you can request that your model include main effects only, effects up to two-way interactions, or effects up to three-way interactions.

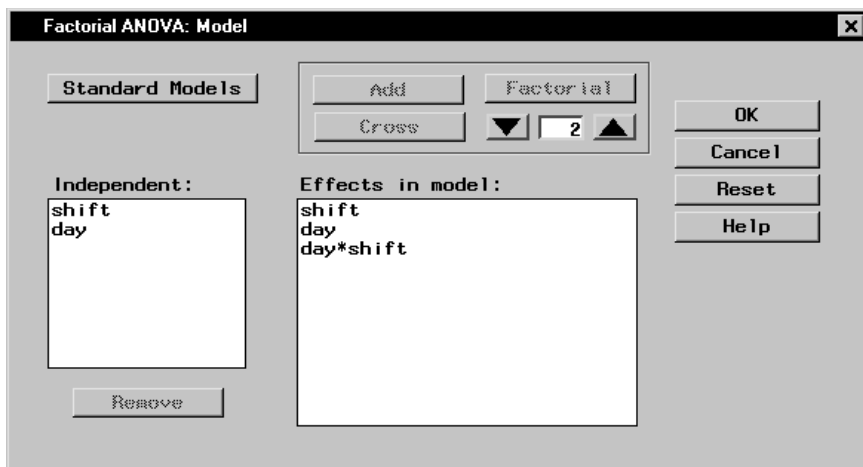


Figure 10.15. Factorial ANOVA: Model Dialog

### Request a Means Plot

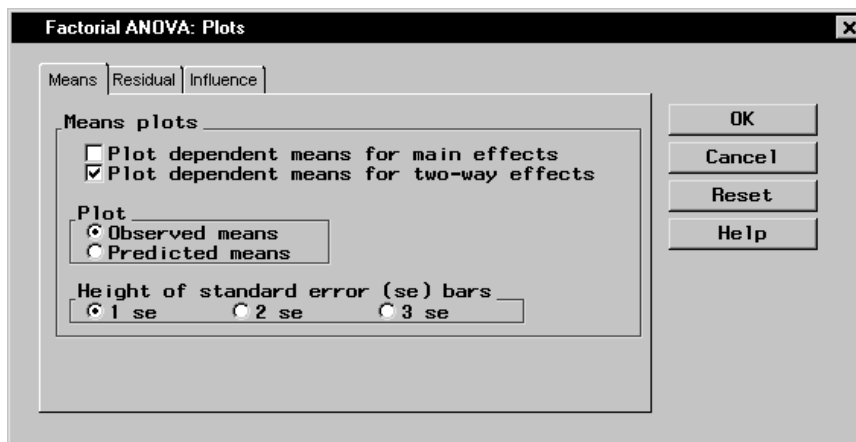
A means plot displays a symbol for the observed or predicted means at each level of a specified variable, with vertical bars extending for a specified number of standard errors. The means for each level of an effect are joined with line segments. To request a plot of the dependent means, follow these steps:

1. Click on the **Plots** button in the main dialog. The resulting window displays the **Means** tab.
2. Select **Plot dependent means for two-way effects**.

You can choose to plot either the observed or predicted means of the dependent variable. Additionally, you can choose whether the vertical bars should represent one, two, or three standard errors.

3. Click **OK**.

Figure 10.16 requests a plot of the observed dependent means for the two-way effects.



**Figure 10.16.** Factorial ANOVA: Plots Dialog

Click **OK** in the main dialog to perform the analysis.

### Review the Results

Figure 10.17 displays information on the levels of the two classification variables, shift and day, followed by the ANOVA table. The model sum of squares is partitioned into the separate contributions of the individual model effects, and  $F$  tests are provided for each effect.

The GLM Procedure						
Class Level Information						
Class	Levels	Values				
shift	3	daytime early late				
day	7	Fri Mon Sat Sun Thu Tue Wed				
Number of observations						168
The GLM Procedure						
Dependent Variable: o3 Ozone						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	20	1526.938137	76.346907	15.37	<.0001	
Error	147	730.355462	4.968405			
Corrected Total	167	2257.293599				
R-Square      Coeff Var      Root MSE      o3 Mean						
0.676446      52.29233      2.228992      4.262560						
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
shift	2	629.9298726	314.9649363	63.39	<.0001	
day	6	347.5540369	57.9256728	11.66	<.0001	
shift*day	12	549.4542274	45.7878523	9.22	<.0001	

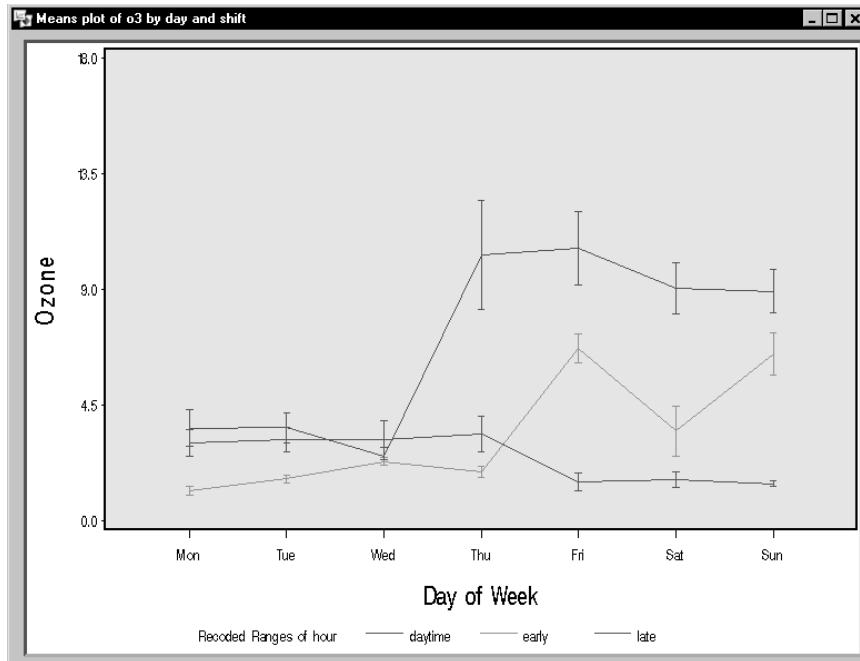
**Figure 10.17.** Factorial ANOVA: Analysis Results

The  $F$  statistic of 15.37 indicates that the model as a whole is highly significant (the  $p$ -value is less than 0.0001). Additionally, the R-square value of 0.6764 means that about 68% of the variation of ozone can be accounted for by the factorial model.

The table at the bottom of Figure 10.17 displays the significance test for each term of the model. The main effects and the interaction term are each significant at the  $\alpha = 0.05$  level (that is, each  $p$ -value is much less than 0.05).



In Figure 10.18, the three curves display ozone concentration across days of the week. Each curve represents the relationship for one of the three factory workshift periods.



**Figure 10.18.** Factorial ANOVA: Means Plot

The means plot indicates an inverse relationship between the daytime and late shifts. The ozone levels during the daytime shift rise dramatically on Thursday and remain high throughout the weekend. Ozone levels for the late shift, on the other hand, start to decrease after Thursday and remain low throughout the weekend.

---

## Linear Models

The Linear Models task enables you to perform an analysis of variance when you have a continuous dependent variable with classification variables, quantitative variables, or both.

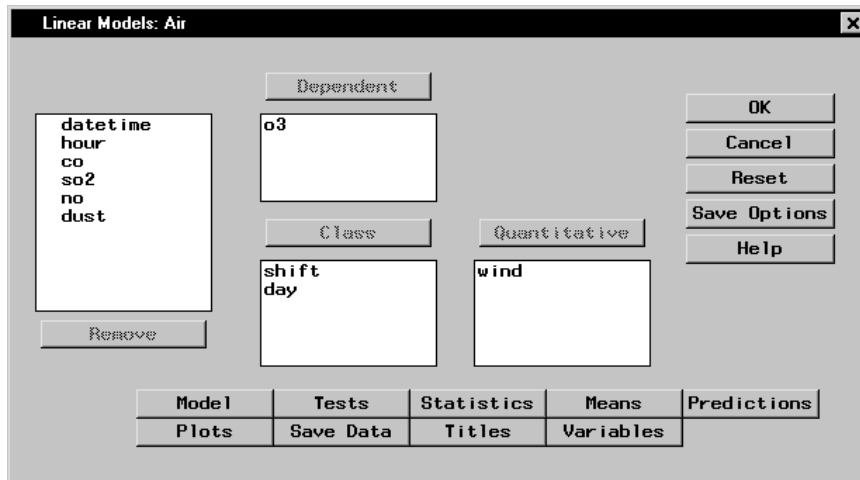
The data set *Air*, described in the section “[The Air Quality Data Set](#)” on page 270, includes quantitative measures; for example, the variable *wind* represents wind speed, in knots. Suppose that you want to model ozone levels using the variables *day* (day of week), *shift* (factory workshift period), and *wind* (wind speed). Suppose that you also want your model to include the interaction between the variables *day* and *shift*. That is, you want to perform a simple two-way analysis of covariance with unequal slopes.

The following example fits this linear model and additionally requests a retrospective power analysis and a plot of the observed values versus the predicted values.

### ***Request the Linear Models Analysis***

To request the linear models analysis, follow these steps:

1. Select **Statistics** → **ANOVA** → **Linear Models . . .**
2. Select *o3* as the dependent variable.
3. Select *shift* and *day* as the class variables.
4. Select *wind* as the quantitative variable.



**Figure 10.19.** Linear Models Dialog

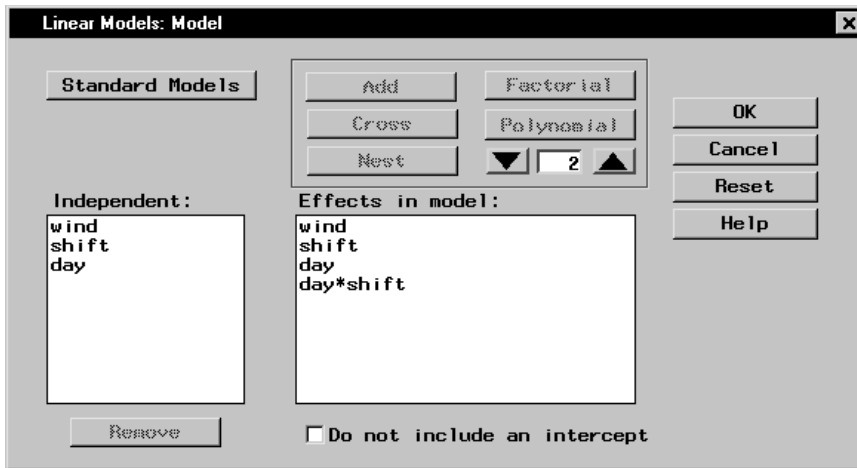
Figure 10.19 displays the Linear Models dialog. By default, the linear model analysis includes only the main effects specified in the main dialog: no interaction term is included.

### ***Specifying an Interaction Term in the Model***

To include the interaction term  $\text{shift}*\text{day}$  in your model, follow these steps:

1. Click on the **Model** button in the main dialog.
2. Highlight the variables **shift** and **day**.
3. Click on the **Cross** button.
4. Click **OK**.

Note that you can build specific models with the **Add**, **Cross**, and **Factorial** buttons, or you can select a model by clicking on the **Standard Models** button and making a selection from the pop-up list.



**Figure 10.20.** Linear Models: Model Dialog

Figure 10.20 displays the Model dialog with the terms `shift` and `day` and the interaction term `shift*day` selected as effects in the model.

### **Request a Power Analysis**

The power of a test is the probability of correctly rejecting the null hypothesis of no difference. It depends on the sample size as well as the precise difference specified in the alternative hypothesis. Ideally, you consider power before gathering data to ensure that you gather enough data to detect a difference. However, once you have gathered your data, you can perform a retrospective power analysis in order to determine how much data is needed to detect the observed difference. To perform a retrospective power analysis with the Analyst Application, follow these steps:

1. Click on the **Tests** button in the main dialog.
2. Click on the **Power Analysis** tab.
3. Select **Perform power analysis**.

To request power calculations for tests performed at several  $\alpha$  values, you can enter the values, separated by a space, in the box labeled **Alphas**. You

can request power analysis for additional sample sizes in the **Sample sizes** box. You can enter one or more specific values for the sample sizes, or you can specify a series of sample sizes in the boxes labeled **From:**, **To:**, and **By:**.

4. Click **OK**.

Figure 10.21 displays the **Power Analysis** tab, which requests a retrospective power analysis with an alpha, or significance level, of 0.05.

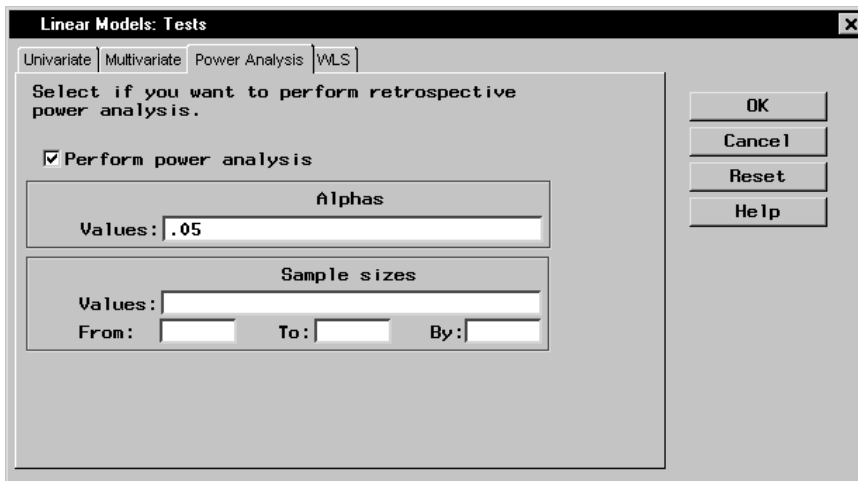
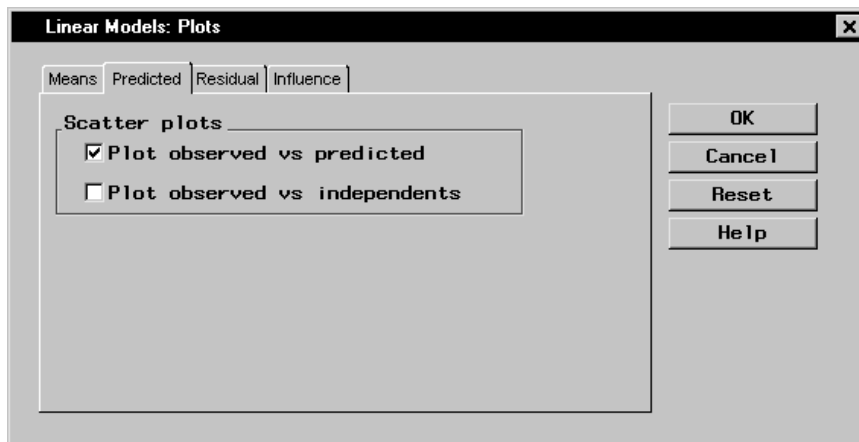


Figure 10.21. Linear Models: Tests Dialog

### **Request a Scatter Plot**

To request a scatter plot of the predicted values versus the observed values, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Click on the **Predicted** tab.
3. Select **Plot observed vs predicted**.
4. Click **OK**.



**Figure 10.22.** Linear Models: Plots Dialog

Figure 10.22 displays the **Predicted** tab in the Plots dialog.

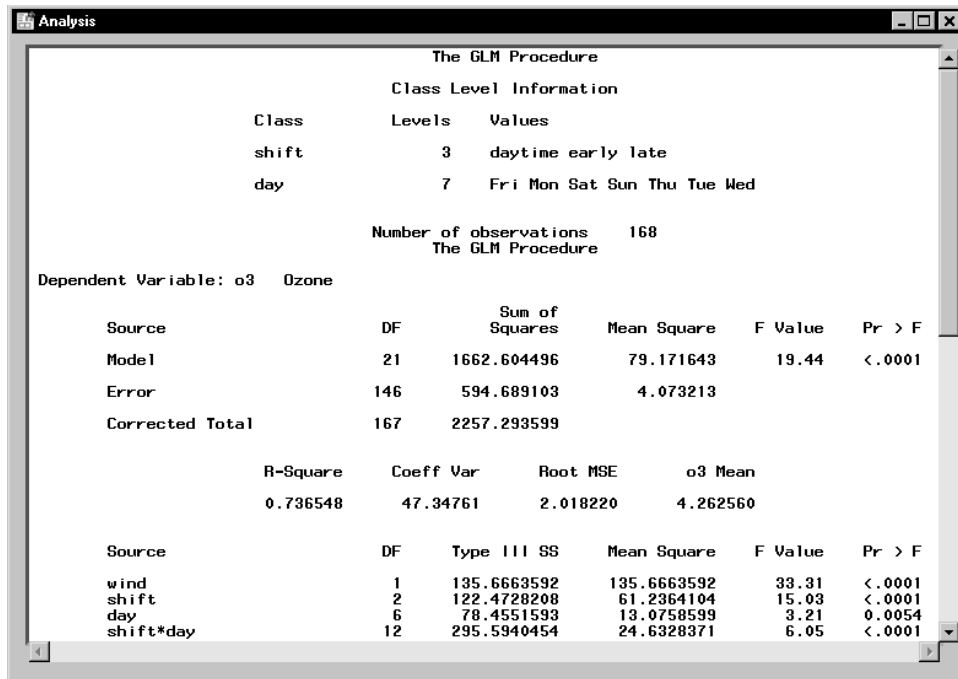
Click **OK** in the Linear Models dialog to perform the analysis.

### **Review the Results**

The output of the analysis includes information about the levels of the independent variables, followed by the ANOVA table.

Figure 10.23 displays the analysis of variance table, with an  $F$  statistic of 19.44 and an associated  $p$ -value less than 0.0001. A  $p$ -value this small indicates that the model explains a highly significant proportion of the variation in the dependent variable.

The R-square value represents the proportion of variability accounted for by the independent variables. In this analysis, about 74% of the variation of the ozone level can be accounted for by the model (that is, by mean differences in day and shift, in conjunction with a linear dependence on wind speed).



The GLM Procedure

Class Level Information

Class	Levels	Values
shift	3	daytime early late
day	7	Fri Mon Sat Sun Thu Tue Wed

Number of observations 168  
The GLM Procedure

Dependent Variable: o3 Ozone

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	1662.604496	79.171643	19.44	<.0001
Error	146	594.689103	4.073213		
Corrected Total	167	2257.293599			

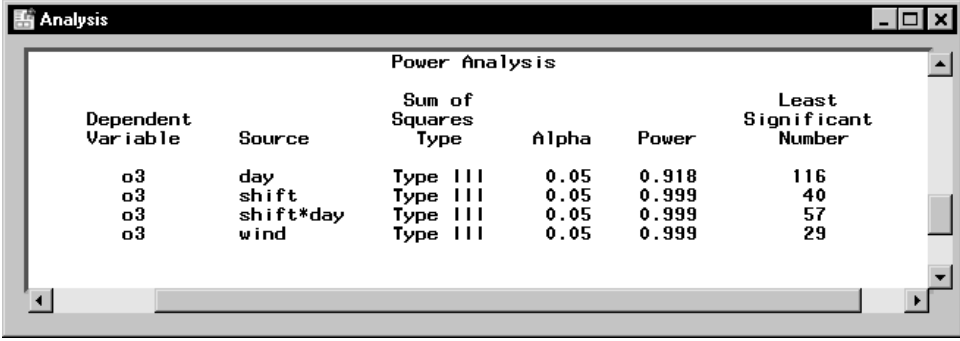
	R-Square	Coeff Var	Root MSE	o3 Mean
	0.736548	47.34761	2.018220	4.262560

Source	DF	Type III SS	Mean Square	F Value	Pr > F
wind	1	135.6663592	135.6663592	33.31	<.0001
shift	2	122.4728208	61.2364104	15.03	<.0001
day	6	78.4551593	13.0758599	3.21	0.0054
shift*day	12	295.5940454	24.6328371	6.05	<.0001

**Figure 10.23.** Linear Models: ANOVA Results

The last table displayed in [Figure 10.23](#) partitions the model sum of squares into the separate contribution for each model effect and tests for the significance of each effect. The main effects and the interaction term are significant at the  $\alpha = 0.05$  level (that is, each  $p$ -value is less than 0.05).

[Figure 10.24](#) displays the retrospective power analysis. The observed power is given for each effect in the linear model.



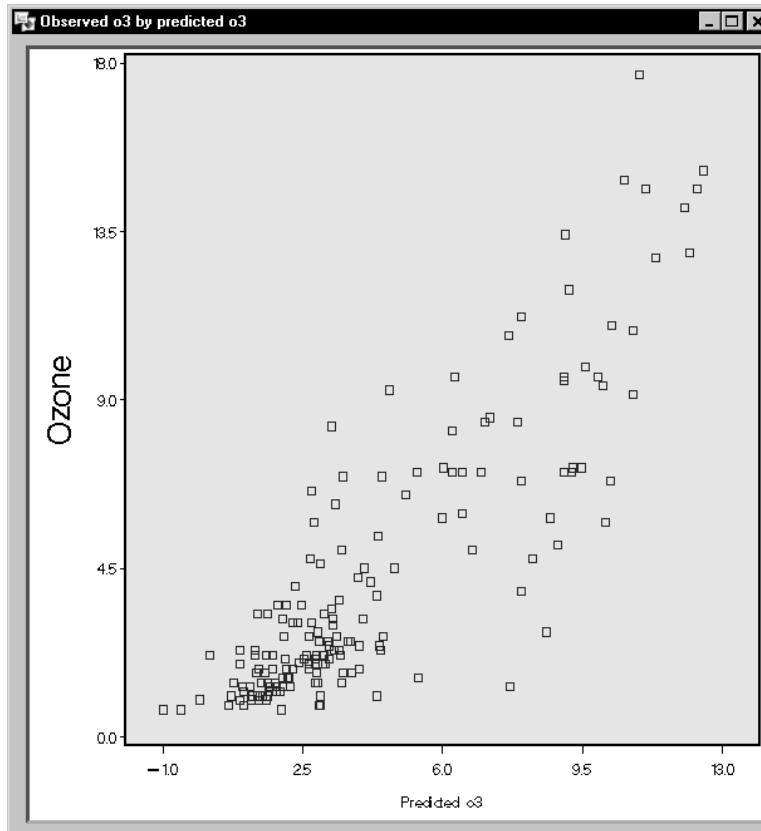
The screenshot shows a window titled "Analysis" containing a table titled "Power Analysis". The table has six columns: "Dependent Variable", "Source", "Sum of Squares Type", "Alpha", "Power", and "Least Significant Number". The data is as follows:

Dependent Variable	Source	Sum of Squares Type	Alpha	Power	Least Significant Number
o3	day	Type III	0.05	0.918	116
o3	shift	Type III	0.05	0.999	40
o3	shift*day	Type III	0.05	0.999	57
o3	wind	Type III	0.05	0.999	29

**Figure 10.24.** Linear Models: Power Analysis

The column labeled Least Significant Number in [Figure 10.24](#) displays the smallest number of observations required to determine that the effect is significant at the given  $\alpha$  value.





**Figure 10.25.** Linear Models: Observed Ozone Levels versus Predicted Values

Figure 10.25 displays the plot of the observed values versus the predicted values from the model. If the model predicts the observed values perfectly, the points on the plot fall on a straight line with a slope of 1. This plot indicates reasonable prediction.

---

## References

SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.

Littell, Ramon C., Freund, Rudolf J., and Spector, Philip C. (1991), *SAS System for Linear Models, Third Edition* by Ramon C. Littell, Rudolf J. Freund, and Philip C. Spector

# Chapter 11

# Regression

## Chapter Contents

---

<b>Introduction</b> . . . . .	301
<b>Simple Linear Regression</b> . . . . .	302
<b>Multiple Linear Regression</b> . . . . .	307
<b>Logistic Regression</b> . . . . .	316
<b>References</b> . . . . .	323



# Chapter 11

## Regression

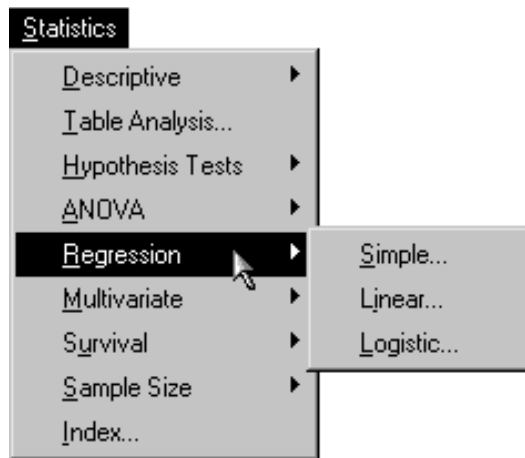
---

### Introduction

Regression techniques enable you to investigate the relationship between a dependent variable (also called a *response* variable) and one or more explanatory variables (also called *predictor*, or *independent*, variables). In linear regression, the dependent variable is modeled as a linear function of the quantitative independent variables. For example, you can write the simple linear regression equation as

$$Y = b_0 + b_1X$$

where  $Y$  represents the single dependent variable,  $X$  is the explanatory variable, and  $b_0$  and  $b_1$  are regression coefficients.



**Figure 11.1.** Regression Menu

The Analyst Application enables you to perform simple linear regression, multiple linear regression and logistic regression. In the Simple linear re-

gression task, you model your dependent variable using a single explanatory variable. In the Linear regression task, you model your dependent variable using one or more explanatory variables. In the Logistic regression task, the dependent variable is discrete, and you model the variable using one or more explanatory variables.

The examples in this chapter demonstrate how you can use the Analyst Application to perform simple linear regression, multiple linear regression, and logistic regression.

---

## Simple Linear Regression

In simple linear regression, there is a single quantitative independent variable. Suppose, for example, that you want to determine whether a linear relationship exists between the asking price for a house and its area in square feet. The area of the house is the quantitative independent variable, and the asking price for the house is the dependent variable.

The data set analyzed in this example is called **Houses**, and it contains the characteristics of fifteen houses for sale. The data set contains the following variables.

<code>style</code>	style category (ranch, split-level, condominium, or two-story)
<code>sqfeet</code>	area in square feet
<code>bedrooms</code>	number of bedrooms
<code>baths</code>	number of bathrooms
<code>street</code>	name of the street on which the house is located
<code>price</code>	asking price for the house

The task includes performing a simple regression analysis to predict the variable `price` from the explanatory variable, `sqfeet`.

### ***Open the Houses Data Set***

The data are provided in the Analyst Sample Library. To open the Houses data set, follow these steps:

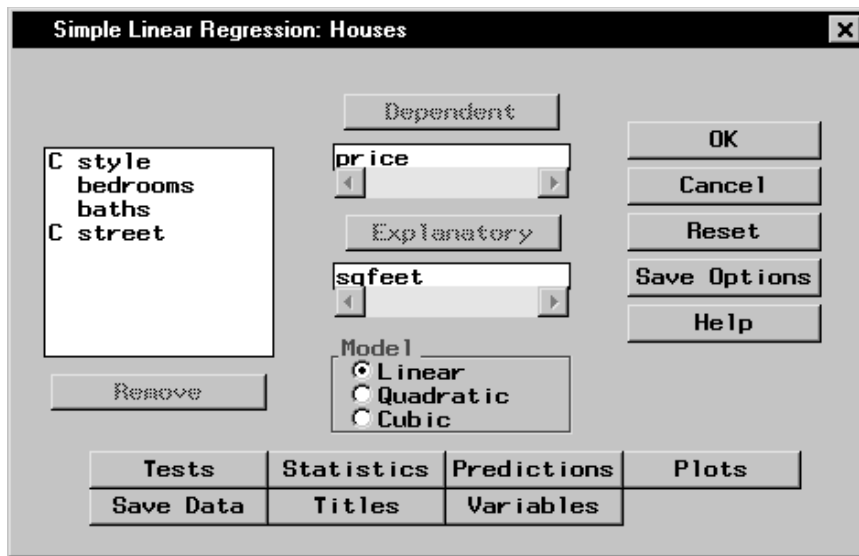
1. Select **Tools** → **Sample Data** . . .
2. Select Houses.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select Houses from the list of members.
7. Click **OK** to bring the Houses data set into the data table.

### ***Request the Simple Regression Analysis***

To request the simple regression analysis, follow these steps:

1. Select **Statistics** → **Regression** → **Simple** . . .
2. Select **price** from the candidate list as the Dependent variable.
3. Select **sqfeet** from the candidate list as the Explanatory variable.

Figure 11.2 displays the resulting dialog.



**Figure 11.2.** Simple Linear Regression Dialog

The model defined in this analysis is

$$\text{price} = b_0 + b_1 \text{sqfeet}$$

If you select **Quadratic** or **Cubic** in the **Model** box, the respective model is

$$\text{price} = b_0 + b_1 \text{sqfeet} + b_2 \text{sqfeet}^2$$

or

$$\text{price} = b_0 + b_1 \text{sqfeet} + b_2 \text{sqfeet}^2 + b_3 \text{sqfeet}^3$$

The default analysis fits the simple regression model.



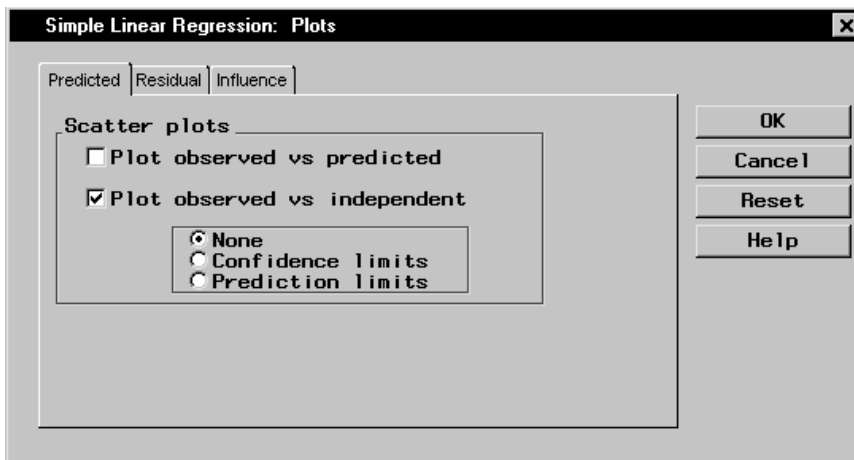
### Request a Scatter Plot of the Data

To request a plot of the observed values versus the independent values, follow these steps.

1. Click on the **Plots** button.
2. Select **Plot observed vs independent**.

You can add 95% confidence limits for the mean of the independent variable by selecting **Confidence limits**, or you can produce 95% prediction limits for individual predictions.

3. Click **OK**.



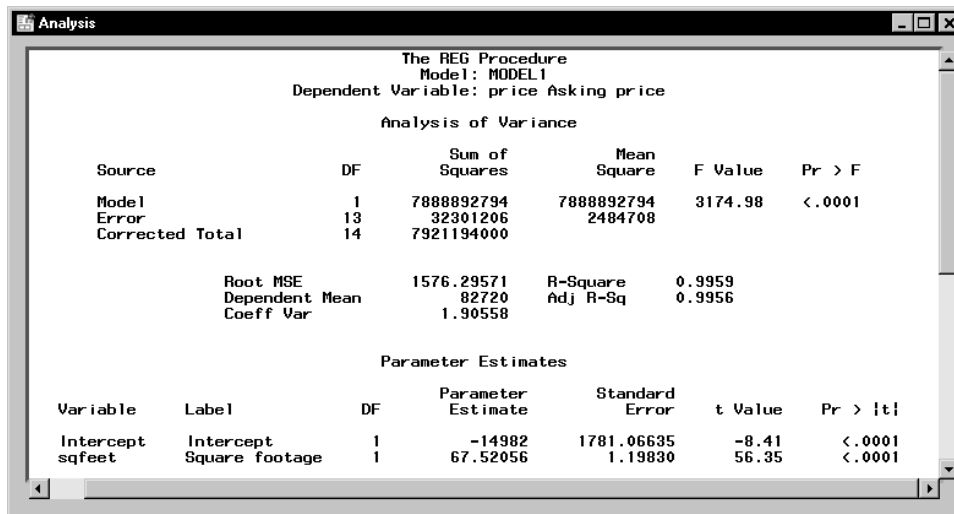
**Figure 11.3.** Simple Linear Regression: Plots Dialog

Click **OK** in the Simple Linear Regression dialog to perform the analysis.

**Review the Results**

The results are displayed in Figure 11.4. The ANOVA table is displayed in the results, followed by the table of parameter estimates. The least squares fit is

$$\text{price} = -14982 + 67.52 \times \text{sqfeet}$$



**Figure 11.4.** Simple Linear Regression: Results

The small  $p$ -values listed in the  $\text{Pr} > |t|$  column indicate that both parameter estimates are significantly different from zero.

The plot of the observed and independent variables is displayed in Figure 11.5. The plot includes the fitted regression line.

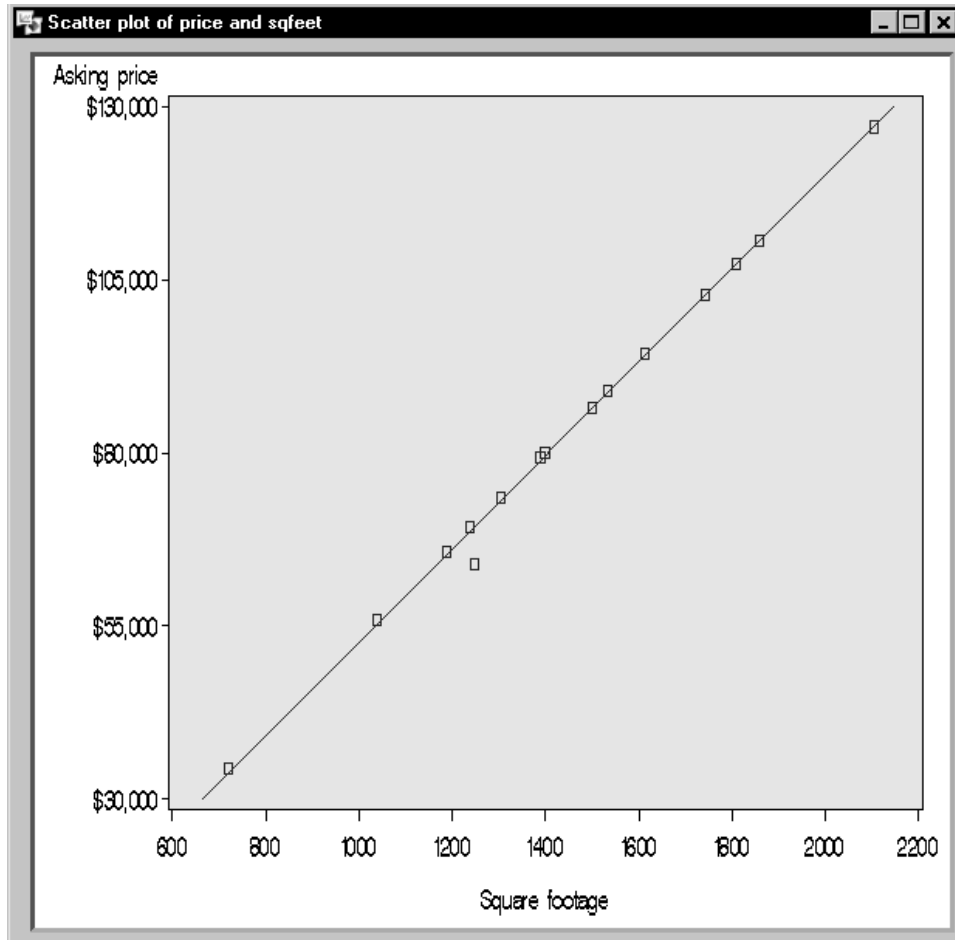


Figure 11.5. Simple Linear Regression: Scatter Plot with Regression Line

---

## Multiple Linear Regression

You perform a multiple linear regression analysis when you have more than one explanatory variable for consideration in your model. You can write the multiple linear regression equation for a model with  $p$  explanatory variables

as

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

where  $Y$  is the response, or dependent, variable, the  $X$ s represent the  $p$  explanatory variables, and the  $b$ s are the regression coefficients.

For example, suppose that you would like to model a person's aerobic fitness as measured by the ability to consume oxygen. The data set analyzed in this example is named **Fitness**, and it contains measurements made on three groups of men involved in a physical fitness course at North Carolina State University. See “[Computing Correlations](#)” in [Chapter 7](#), “[Descriptive Statistics](#),” for a complete description of the variables in the Fitness data set.

The goal of the study is to predict fitness as measured by oxygen consumption. Thus, the dependent variable for the analysis is the variable **oxygen**. You can choose any of the other quantitative variables (**age**, **weight**, **runtime**, **rstpulse**, **runpulse**, and **maxpulse**) as your explanatory variables.

Suppose that previous studies indicate that oxygen consumption is dependent upon the subject's age, the time it takes to run 1.5 miles, and the heart rate while running. Thus, in order to predict oxygen consumption, you estimate the parameters in the following multiple linear regression equation:

$$\text{oxygen} = b_0 + b_1 \text{age} + b_2 \text{runtime} + b_3 \text{runpulse}$$

This task includes performing a linear regression analysis to predict the variable **oxygen** from the explanatory variables **age**, **runtime**, and **runpulse**. Additionally, the task requests confidence intervals for the estimates, a collinearity analysis, and a scatter plot of the residuals.

### **Open the Fitness Data Set**

The data are provided in the Analyst Sample Library. To access this data set, follow these steps:

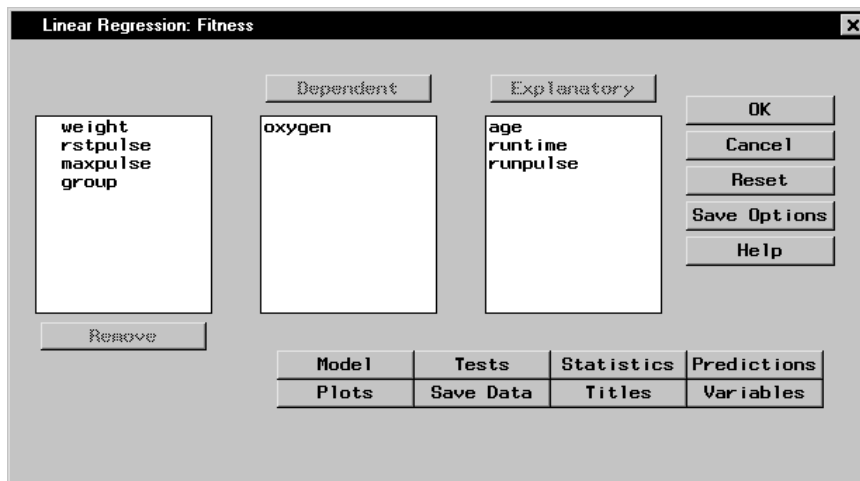
1. Select **Tools** → **Sample Data** . . .
2. Select **Fitness**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Fitness** from the list of members.
7. Click **OK** to bring the **Fitness** data set into the data table.

### **Request the Linear Regression Analysis**

To specify the analysis, follow these steps:

1. Select **Statistics** → **Regression** → **Linear** . . .
2. Select the variable **oxygen** from the candidate list as the dependent variable.
3. Select the variables **age**, **runtime**, and **runpulse** as the explanatory variables.

Figure 11.6 displays the resulting Linear Regression task.



**Figure 11.6.** Linear Regression Dialog

The default analysis fits the linear regression model.

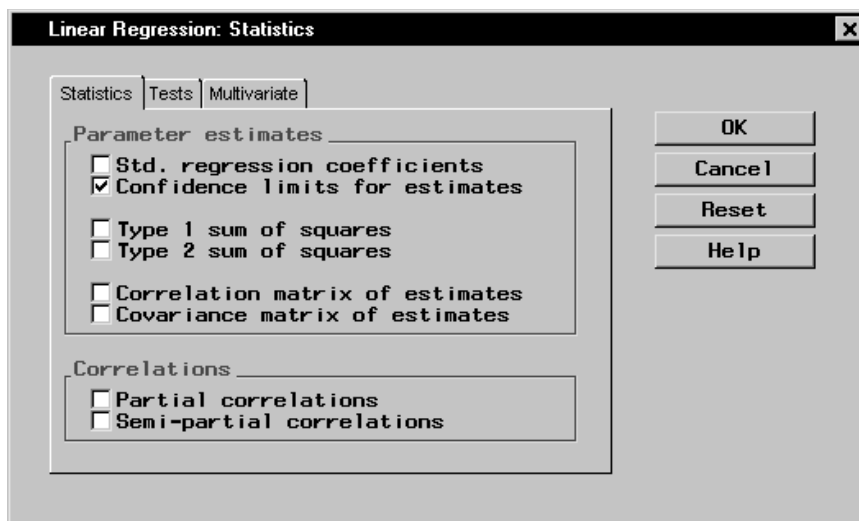
### Request Additional Statistics

You can request several additional statistics for your analysis in the Statistics dialog.

To request that confidence limits be computed, follow these steps:

1. Click on the **Statistics** button.
2. In the **Statistics** tab, select **Confidence limits for estimates**.

Figure 11.7 displays the **Statistics** tab in the Statistics dialog.

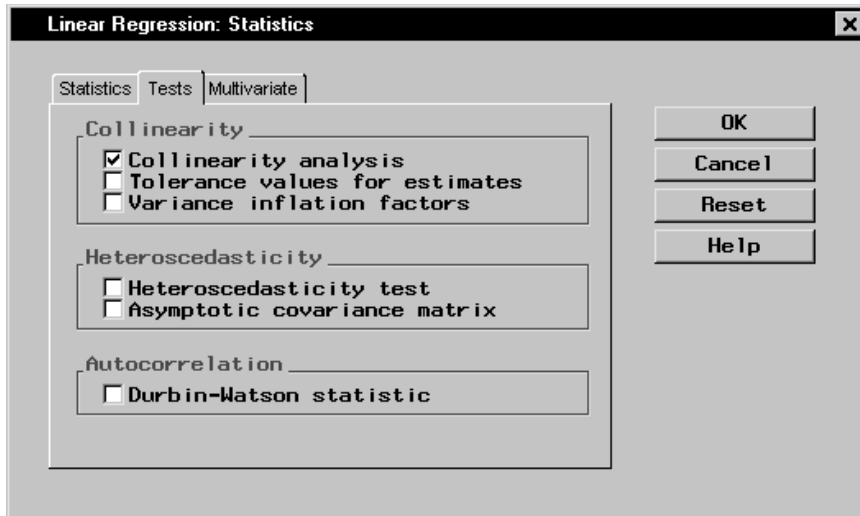


**Figure 11.7.** Linear Regression: Statistics Dialog, Statistics Tab

To request a collinearity analysis, follow these steps:

1. Click on the **Tests** tab in the Statistics dialog.
2. Select **Collinearity analysis**.
3. Click **OK**.

The dialog in [Figure 11.8](#) requests a collinearity analysis in order to assess dependencies among the explanatory variables.



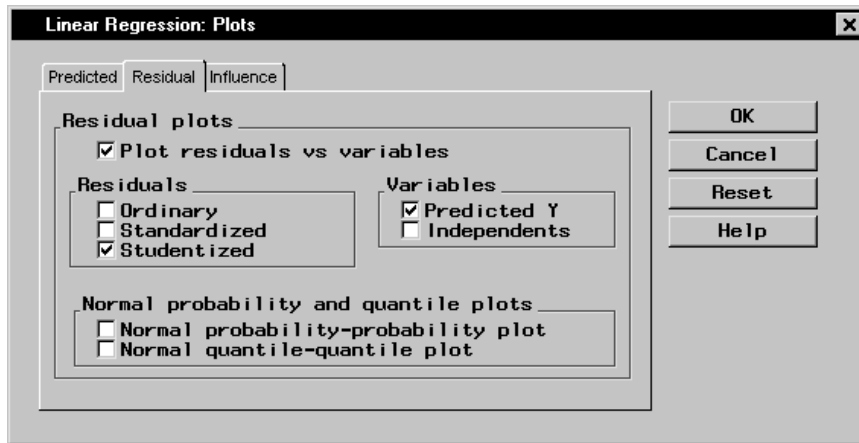
**Figure 11.8.** Linear Regression: Statistics Dialog, Tests Tab

### ***Request a Scatter Plot of the Residuals***

To request a plot of the studentized residuals versus the predicted values, follow these steps:

1. In the Linear Regression main dialog, click on the **Plots** button.
2. Click on the **Residual** tab.
3. Select **Plot residuals vs variables**.
4. In the box labeled **Residuals**, check the selection **Studentized**.
5. In the box labeled **Variables**, check the selection **Predicted Y**.
6. Click **OK**.

[Figure 11.9](#) displays the **Residual** tab.



**Figure 11.9.** Linear Regression: Plots Dialog, Residual Tab

An ordinary residual is the difference between the observed response and the predicted value for that response. The standardized residual is the ratio of the residual to its standard error; that is, it is the ordinary residual divided by its standard error. The studentized residual is the standardized residual calculated with the current observation deleted from the analysis.

Click **OK** in the Linear Regression dialog to perform the analysis.

### **Review the Results**

Figure 11.10 displays the analysis of variance table and the parameter estimates.



The REG Procedure  
Model: MODEL1  
Dependent Variable: oxygen Oxygen consumption

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	690.55086	230.18362	38.64	<.0001
Error	27	160.83069	5.95669		
Corrected Total	30	851.38154			

Root MSE	2.44063	R-Square	0.8111
Dependent Mean	47.37581	Adj R-Sq	0.7901
Coeff Var	5.15165		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	111.71806	10.23509	10.92	<.0001
age	Age in years	1	-0.25640	0.09623	-2.66	0.0129
runtime	Min. to run 1.5 miles	1	-2.82538	0.35828	-7.89	<.0001
runpulse	Heart rate while running	1	-0.13091	0.05059	-2.59	0.0154

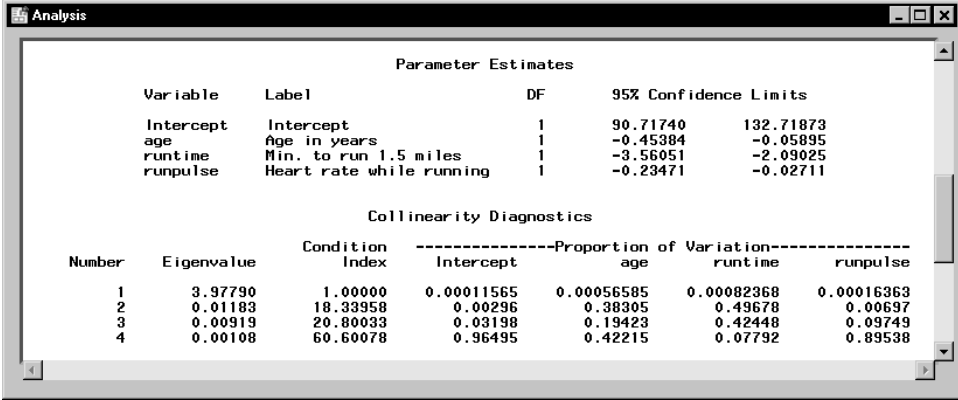
**Figure 11.10.** Linear Regression: ANOVA Table and Parameter Estimates

In the analysis of variance table displayed in [Figure 11.10](#), the  $F$  value of 38.64 (with an associated  $p$ -value that is less than 0.0001) indicates a significant relationship between the dependent variable, **oxygen**, and at least one of the explanatory variables. The R-square value indicates that the model accounts for 81% of the variation in oxygen consumption.

The “Parameter Estimates” table lists the degrees of freedom, the parameter estimates, and the standard error of the estimates. The final two columns of the table provide the calculated  $t$  values and associated probabilities ( $p$ -values) of obtaining a larger absolute  $t$  value. Each  $p$ -value is less than 0.05; thus, all parameter estimates are significant at the 5% level. The fitted equation for this model is as follows:

$$\text{oxygen} = 111.718 - 0.256 \times \text{age} - 2.825 \times \text{runtime} - 0.131 \times \text{runpulse}$$

[Figure 11.11](#) displays the confidence limits for the parameter estimates and the table of collinearity diagnostics.



The screenshot shows a software window titled "Analysis" with two tables. The first table, "Parameter Estimates", lists variables, their labels, degrees of freedom (DF), and 95% confidence limits. The second table, "Collinearity Diagnostics", lists the number of components, eigenvalues, condition indices, and the proportion of variation accounted for by each component across the four variables: Intercept, age, runtime, and runpulse.

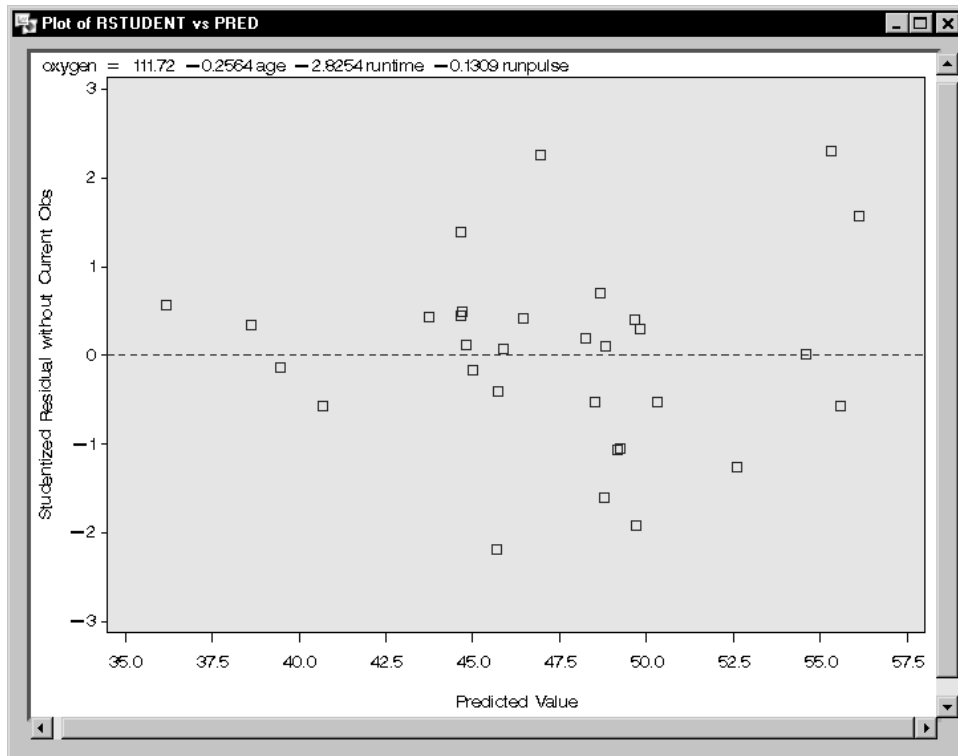
Parameter Estimates					
Variable	Label	DF	95% Confidence Limits		
Intercept	Intercept	1	90.71740	132.71873	
age	Age in years	1	-0.45384	-0.05895	
runtime	Min. to run 1.5 miles	1	-3.56051	-2.09025	
runpulse	Heart rate while running	1	-0.23471	-0.02711	

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	-----Proportion of Variation-----			
			Intercept	age	runtime	runpulse
1	3.97790	1.00000	0.00011565	0.00056585	0.00082368	0.00016363
2	0.01183	18.33958	0.00296	0.38305	0.49678	0.00697
3	0.00919	20.80033	0.03198	0.19423	0.42448	0.09749
4	0.00108	60.60078	0.96495	0.42215	0.07792	0.89538

**Figure 11.11.** Linear Regression: Confidence Limits and Collinearity Analysis

The collinearity diagnostics table displays the eigenvalues, the condition index, and the corresponding proportion of variation accounted for in each estimate. Generally, when the condition index is around 10, there are weak dependencies among the regression estimates. When the index is larger than 100, the estimates may have a large amount of numerical error. The diagnostics displayed in [Figure 11.11](#), though indicating unfavorable dependencies among the estimates, are not so excessive as to dismiss the model.



**Figure 11.12.** Linear Regression: Plot of Studentized Residuals versus Predicted Values

The plot of the studentized residuals versus the predicted values is displayed in [Figure 11.12](#). When a model provides a good fit and does not violate any model assumptions, this type of residual plot exhibits no marked pattern or trend. [Figure 11.12](#) exhibits no such trend, indicating an adequate fit.

## Logistic Regression

Logistic regression enables you to investigate the relationship between a categorical outcome and a set of explanatory variables. The outcome, or response, can be dichotomous (yes, no) or ordinal (low, medium, high). When you have a dichotomous response, you are performing standard logistic regression. When you are modeling an ordinal response, you are fitting a proportional odds model.

You can express the logistic model for describing the variation among probabilities  $\{\theta_h\}$  as

$$\theta_h = \{1 + \exp[-\alpha - \sum_{k=1}^t \beta_k x_{hk}]\}^{-1}$$

where  $\alpha$  is the intercept parameter,  $\beta$  is a vector of  $t$  regression parameters, and  $\mathbf{x}'_h$  is a row vector of explanatory variables corresponding to the  $h$ th subpopulation.

You can show that the odds of success for the  $h$ th group are

$$\frac{\theta_h}{1 - \theta_h} = \exp\{\alpha + \sum_{k=1}^t \beta_k x_{hk}\}$$

By taking logs on both sides, you obtain a linear model for the *logit*:

$$\log\left\{\frac{\theta_h}{1 - \theta_h}\right\} = \alpha + \sum_{k=1}^t \beta_k x_{hk}$$

This is the log odds of success to failure for the  $h$ th subpopulation. A nice property of the logistic model is that all possible values of  $(\alpha + \mathbf{x}'_h\beta)$  in  $(-\infty, \infty)$  map into  $(0, 1)$  for  $\theta_h$ . Note that  $\exp\{\beta_k\}$  are the odds ratios. Maximum likelihood methods are used to estimate  $\alpha$  and  $\beta$ .

In a study on the presence of coronary artery disease, walk-in patients at a clinic were examined for symptoms of coronary artery disease. Investigators also administered an ECG. Interest lies in determining whether there is a relationship between presence or absence of coronary artery disease and ECG score and gender of patient. Logistic regression is the appropriate tool for such an investigation.

The data set analyzed in this example is called **Coronary2**. It contains the following variables:

sex	sex (m or f)
ecg	ST segment depression (low, medium, or high)
age	patient age
ca	disease (yes or no)

The task includes performing a logistic analysis to determine an appropriate model.

### **Open the Coronary2 Data Set**

The data are provided in the Analyst Sample Library. To open the Coronary2 data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Coronary2.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Coronary2 from the list of members.
7. Click **OK** to bring the Coronary2 data set into the data table.

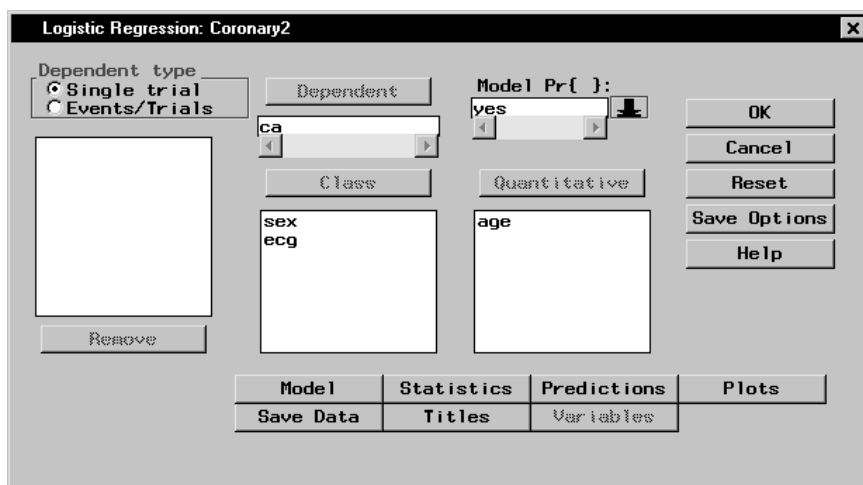
### **Request the Logistic Regression Analysis**

To request the logistic regression analysis, follow these steps:

1. Select **Statistics** → **Regression** → **Logistic ...**
2. Ensure that **Single trial** is selected as the **Dependent type**.
3. Select **ca** from the candidate list as the dependent variable.
4. Select **ecg** and **sex** from the candidate list as the class variables.
5. Select **age** from the candidate list as the quantitative variable.
6. Select **yes** from the drop-down list for **Model Pr{ }:**

Note that **Model Pr{ }:** determines which value of the dependent variable the model is based on; usually, the value representing an event (such as yes or success) is chosen.

Figure 11.13 displays the resulting dialog.



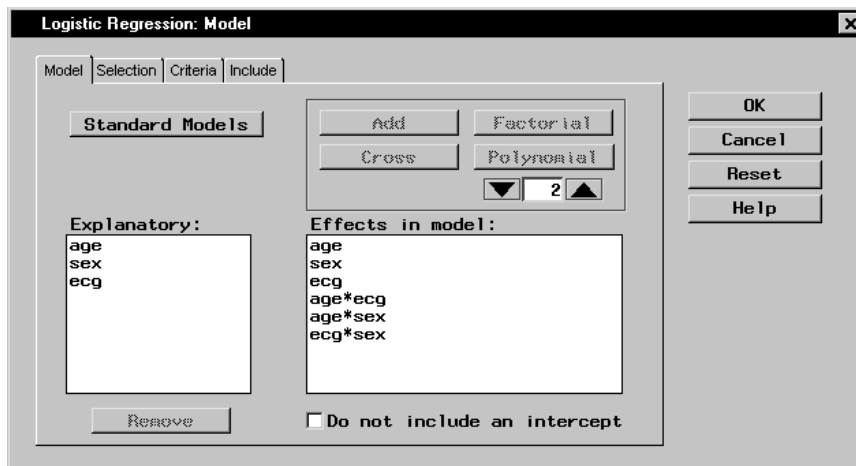
**Figure 11.13.** Logistic Regression Dialog

### **Specify the Model**

By default, a main effects model is fit. To define a different model, with terms such as interactions, or to specify various model selection methods, such as forward selection or backward elimination, use the Model dialog.

To specify a forward selection model with main effects and their interactions, follow these steps:

1. Click on the **Model** button in the main dialog.
2. Highlight the variables **age**, **ecg**, and **sex** in the **Explanatory:** list of the model dialog.
3. Click on the **Factorial** button to specify main effects and their interactions.



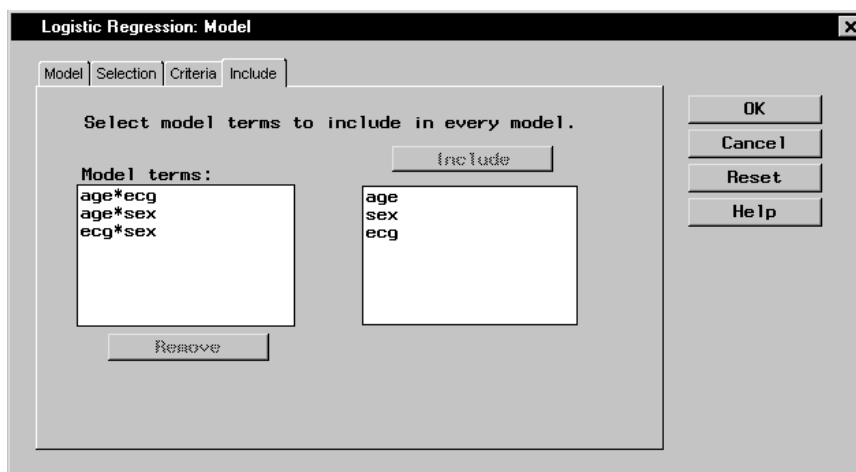
**Figure 11.14.** Logistic Regression: Model Dialog, Model Tab

Figure 11.14 displays the Model dialog with the terms **age**, **ecg**, **sex**, and their interactions selected as effects in the model.

Note that you can build specific models with the **Add**, **Cross**, and **Factorial** buttons, or you can select a model by clicking on the **Standard Models** button and making a selection from the pop-up list. From this list, you can request that your model include main effects only or effects up to two-way interactions.

Now, to specify your model-building technique, follow these steps:

1. Click on the **Selection** tab.
2. Select **Forward selection**. The forward selection technique starts with a default model and adds significant variables to the model according to the specified criteria.
3. To specify which variables to include in every model, click on the **Include** tab, and select the variables **age**, **ecg**, and **sex**.
4. Click **OK**.



**Figure 11.15.** Logistic Regression: Model Dialog, Include Tab

Figure 11.15 displays the **Include** tab with the terms **age**, **ecg**, and **sex** selected as model terms to be included in every model.

When you have completed your selections, click **OK** in the main dialog to produce your analysis.



## Review the Results

Figure 11.16 displays the “Testing Global Null Hypothesis: BETA = 0” table, which lists statistics that test whether the parameters are collectively equal to zero. This is similar to the overall  $F$  statistic in a regression model.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.4878	4	0.0003
Score	18.9094	4	0.0008
Wald	14.6894	4	0.0054

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
2.2464	5	0.8141

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	7.2340	0.0072
sex	1	6.3416	0.0118
ecg	2	5.6706	0.0587

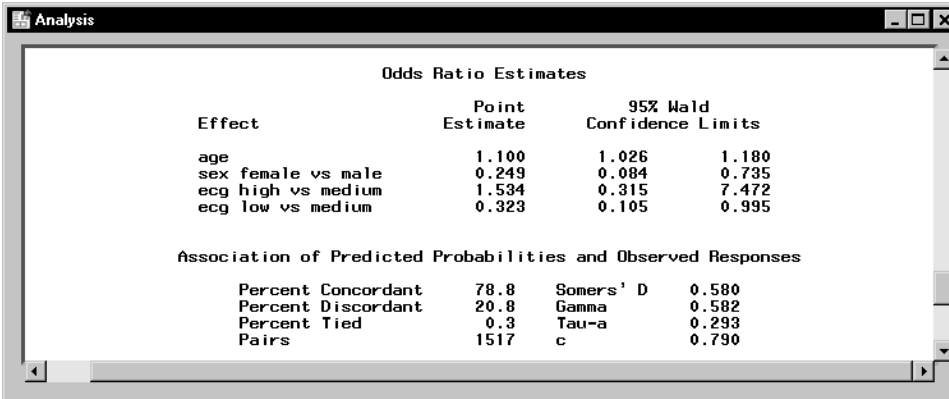
**Figure 11.16.** Logistic Regression: Analysis Results

When the explanatory variables in a logistic regression are relatively small in number and are qualitative, you can request a goodness-of-fit test. However, when you also have quantitative variables, the sample size requirements for these tests are not met. An alternative strategy for testing goodness of fit in this case is to examine the residual score statistic. This criterion is based on the relationship of the residuals of the model with other potential explanatory variables. If an association exists, then the additional explanatory variable should also be included in the model. This test is distributed as chi-square, with degrees of freedom equal to the difference in the number of parameters in the original model and the number of parameters in the expanded model.

The residual score statistic is displayed in Figure 11.16 as the “Residual Chi-Square Test” table. Since the difference in the number of parameters for the

expanded model and the original model is  $9 - 4 = 5$ , the score statistic has 5 degrees of freedom. Since the value of the statistic is 2.24 and the  $p$ -value is 0.81, the main effects model fits adequately and no additional interactions need to be added.

The “Type III Tests of Effects” table provides Wald chi-square statistics that indicate that both **age** and **sex** are clearly significant at the  $\alpha = 0.05$  level of significance. The **ecg** variable approaches significance, with the Wald statistic of 5.67 and  $p = 0.059$ . Although you may want to delete the **ecg** variable because it does not meet the  $\alpha = 0.05$  significance criteria, there may be reasons for keeping it.



The screenshot shows a window titled "Analysis" with two tables of results. The first table, "Odds Ratio Estimates", lists the point estimates and 95% Wald confidence limits for four effects: age, sex female vs male, ecg high vs medium, and ecg low vs medium. The second table, "Association of Predicted Probabilities and Observed Responses", provides summary statistics including Percent Concordant, Percent Discordant, Percent Tied Pairs, Somers' D, Gamma, Tau-a, and c.

Odds Ratio Estimates				
Effect	Point Estimate	95% Wald Confidence Limits		
age	1.100	1.026	1.180	
sex female vs male	0.249	0.084	0.735	
ecg high vs medium	1.534	0.315	7.472	
ecg low vs medium	0.323	0.105	0.995	

Association of Predicted Probabilities and Observed Responses				
Percent Concordant	78.8	Somers' D	0.580	
Percent Discordant	20.8	Gamma	0.582	
Percent Tied	0.3	Tau-a	0.293	
Pairs	1517	c	0.790	

**Figure 11.17.** Logistic Regression: Analysis Results

Figure 11.17 displays odds ratio estimates and statistics describing the association of predicted probabilities and observed responses. The value of 1.10 for **age** is the extent to which the odds of coronary heart disease increase each year. The odds ratio for **sex**, 0.249, is the odds for females relative to males adjusted for **age** and **ecg**. Thus, the odds of coronary heart diseases for females are approximately one-fourth that of males.

---

## References

- Freund, Rudolf J. and Littell, Ramon C. (1991), *SAS System for Regression, Second Edition*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.
- Stokes, Maura E., Davis, Charles S., and Koch, Gary G. (1995), *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc.



# Chapter 12

## Sample Size and Power Calculations

### Chapter Contents

---

<b>Introduction</b> . . . . .	327
<b>Hypothesis Testing</b> . . . . .	329
<b>Confidence Intervals</b> . . . . .	333
<b>Equivalence Tests</b> . . . . .	338
<b>One-Way ANOVA</b> . . . . .	343
<b>Power Computation Details</b> . . . . .	346
Hypothesis Tests . . . . .	346
One-Way ANOVA . . . . .	348
Confidence Intervals . . . . .	348
Equivalence Tests . . . . .	350
<b>References</b> . . . . .	352

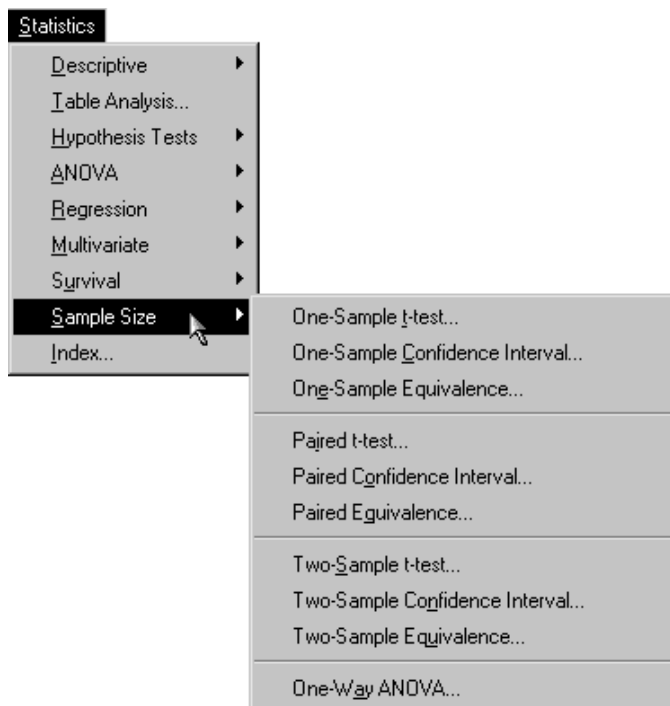


# Chapter 12

## Sample Size and Power Calculations

---

### Introduction



**Figure 12.1.** Sample Size Menu

When you are planning a study or experiment, you often need to know how many units to sample to obtain a certain power, or you may want to know the power you would obtain with a specific sample size. The *power* of a hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is true. With an inadequate sample size, you may not reach valid

conclusions with your work; with an excessive sample size, you may waste valuable resources. Thus, performing sample size and power computations is often quite important.

The power and sample size calculations depend on the planned data analysis strategy. That is, if the primary hypothesis test is a two-sample  $t$ -test, then the power calculations must be based on that test. Otherwise, if the sample size calculations and data analyses are not aligned, the results may not be correct.

Determining sample size requirements ahead of the experiment is a prospective exercise. Then, you proceed to select the appropriate number of sampling units and perform data collection and analysis. However, power and sample size calculations are also useful retrospectively. For a given analysis, you may want to calculate what level of power you achieved or what sample size would have been needed for a given power.

Power and sample size calculations are a function of the specific alternative hypothesis of interest, in addition to other parameters. That is, the power results will vary depending on which value of the alternative hypothesis you specify, so sometimes it is useful to do these analyses for a range of values to see how sensitive the power analysis is to changes in the alternative hypothesis value. Often, you produce plots of power versus sample size, called *power curves*, to see how sample size and power affect each other.

The Sample Size tasks provide prospective sample size and power calculations for several types of analyses:  $t$ -tests, confidence intervals, and tests of equivalence. Each of these calculations is available for one-sample, paired-sample, and two-sample study designs. Power and sample size calculations are also available for the one-way ANOVA design. Multiple parameter values can be input, and results and power curves are produced for each combination of values. Note that retrospective power computations are also available in a number of the statistical tasks in the Analyst Application such as the Hypothesis Test, Regression, and ANOVA tasks.



---

## Hypothesis Testing

Sample size and power calculations are available for one-sample and two-sample paired and independent designs where the proposed analysis is hypothesis testing of a mean or means via a  $t$ -test. These computations assume equally sized groups.

Suppose you want to compute the power for a one-sample  $t$ -test. The alternative hypothesis mean and the standard deviation have the values 8.6137 and 2.0851, respectively. You are interested in testing whether the null mean has the value 8, at an alpha level of 0.05, and you are interested in looking at a range of sample sizes from 11 to 211. The study for which these statistics were computed had a sample size of 51.

### **Requesting Power Computations for the One-Sample $t$ -test**

To access this task, select

**Statistics → Sample Size → One-Sample  $t$ -test . . .**

Figure 12.2 displays the resulting dialog. Note that, unlike the other statistical tasks that require a data set for analysis, performing one of the Sample Size tasks requires only entering information in the appropriate dialog. The data table is not involved.

**Figure 12.2.** Sample Size Dialog for One-Sample t-test

In this task, you specify whether you want to compute sample size or power, enter values for the test hypothesis and parameters, specify the alpha level (0.05 is the default), specify whether you want a power curve produced, and specify a range of power values or sample sizes depending on whether you are computing sample size or power.

To enter the information for this example, follow these steps:

1. Select **Power**.
2. Enter 8 as the **Null mean:** value.
3. Enter 8.6137 as the **Alternate mean:**
4. Enter 2.0851 as the **Standard deviation:**
5. Make sure that the **Alpha:** value is 0.05.
6. Enter 11 as the value for the **From:** field in the line for **N:**
7. Enter 211 and 20 as the values under **To:** and **By:**, respectively, in the line for **N:**
8. Select **Power vs. N** to produce a plot.

9. Enter 51 as the value for **N ref line**:
10. Select **2-sided** for **Tails** if it is not already selected.

Note that you can enter multiple values in fields such as for **Alpha:** and **Null mean:**, separated by commas or blanks, and the analysis will be performed for all combinations of the entered values. Here, power will be computed for sample sizes ranging from 11 to 211 in multiples of 20.

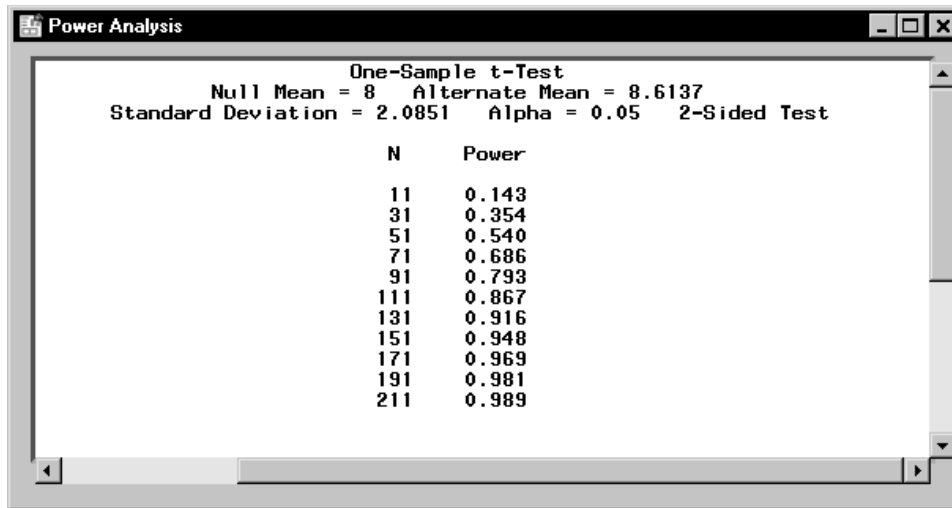
Figure 12.3 contains the completed dialog.

The screenshot shows the 'One-Sample t-test' dialog box with the following settings:

- Calculate:**  Power,  N
- Test specifications:**
  - Null mean: 8
  - Alternate mean: 8.6137
  - Standard deviation: 2.0851
  - Alpha: 0.05
  - N: From: 11, To: 211, By: 20
- Plot:**
  - Power vs. N
  - Power ref line: [ ]
  - N ref line: 51
- Tails:**
  - 1-sided
  - 2-sided
- Buttons:** OK, Cancel, Reset, Save Options, Help, Titles

**Figure 12.3.** Sample Size Dialog for One-Sample t-test

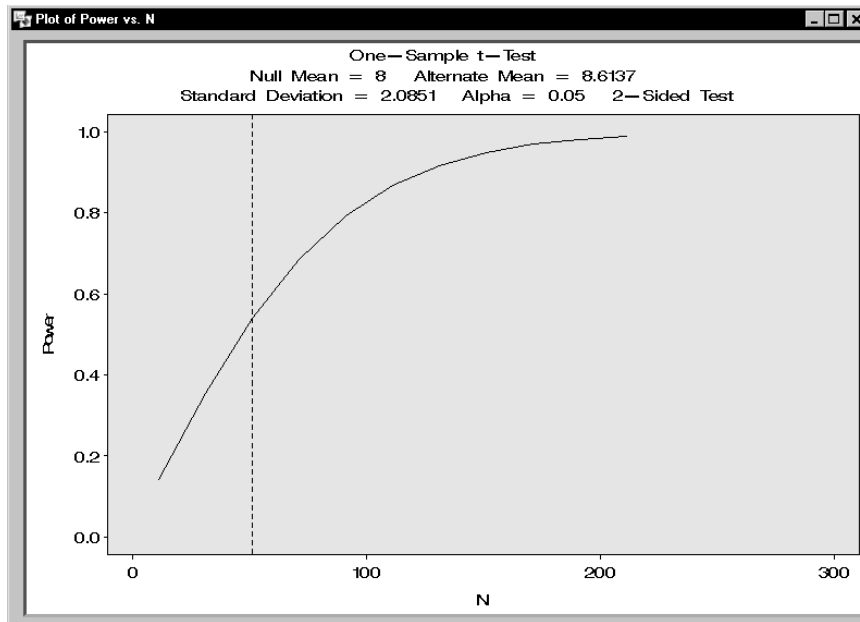
Figure 12.4 contains the power computations for the sample sizes ranging from 11 to 211.



**Figure 12.4.** Sample Size Results for One-Sample t-test

The interpretation of a power of 0.540 for  $n = 51$  is as follows: suppose the true mean and standard deviation are 8.6137 and 2.0851, and suppose a random sample of 51 observations is taken. Then the probability that the hypothesis test will reject the null hypothesis ( $H_0: \mu = 8.0$ ) and conclude (correctly) that the alternative hypothesis ( $H_A: \mu = 8.6137$ ) is true is 0.540.

The requested plot is shown in [Figure 12.5](#) with a reference line at  $n = 51$ .



**Figure 12.5.** Plot of Power versus Sample Size

### More on Hypothesis Tests

In the two-sample cases, you must enter the null means of each group and the standard deviation. In the paired case, the standard deviation entered is the standard deviation of the differences between the two groups. In the independent case, the standard deviation is the pooled standard deviation, which is calculated as follows:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}$$

---

## Confidence Intervals

Sample size and power calculations are available for one-sample and two-sample paired and independent designs, when the proposed analysis is con-

struction of confidence intervals of a mean (one-sample) or difference of two means (two-sample), via the  $t$ -test.

To understand the power of a confidence interval, first define the *precision* to be half the length of a two-sided confidence interval (or the distance between the endpoint and the parameter estimate in a one-sided interval). The power can then be considered to be the probability that the desired precision is achieved, that is, the probability that the length of the two-sided interval is no more than twice the desired precision. Here, a slight modification of this concept is used. The power is considered to be the conditional probability that the desired precision is achieved, given that the interval includes the true value of the parameter of interest. The reason for the modification is that there is no reason for the interval to be particularly small if it does not contain the true value of the parameter.

These computations assume equally sized groups.

### **Requesting Power Computations for a Confidence Interval in a Paired $t$ -test**

To perform this task, select

**Statistics → Sample Size → Paired Confidence Interval . . .**

Figure 12.6 displays the resulting dialog.

**Paired Confidence Interval**

Calculate:  Power  N

Test specifications

Desired precision:

Std dev of diff:

Alpha:

N:  From:  To:  By:

Plot

Power vs. N

Power ref line:

N ref line:

Interval

1-sided

2-sided

OK

Cancel

Reset

Save Options

Help

Titles

**Figure 12.6.** Sample Size Dialog for Paired Confidence Interval

You specify whether you want to compute sample sizes or power, enter values for desired precision and standard deviation, enter the alpha levels, enter the sample sizes or power, and select if you want a power curve.

To request power for a paired confidence interval where the desired precision is 0.5 and the standard deviation is 2.462, follow these steps:

1. Select **Power**.
2. Enter 0.5 as the **Desired precision**:
3. Enter 2.462 as the **Std dev of diff**:
4. Enter 0.01, 0.025, 0.05, and 0.1 as values in the field for **Alpha**:
5. Enter 11 as the value for the **From**: field in the line for **N**:
6. Enter 211 and 5 as the values under **To**: and **By**:, respectively, in the line for **N**:
7. Select **Power vs. N**.
8. Select **2-sided** for **Interval** if it is not already selected.

Note that you can enter multiple values in these fields, for example, for **Alpha:** and **Desired precision:**, separated by commas or blanks, and the analysis will be performed for all combinations of the input values. Here, power will be computed for sample sizes ranging from 11 to 211 in multiples of 5.

**Figure 12.7.** Completed Sample Size Dialog for Paired Confidence Interval

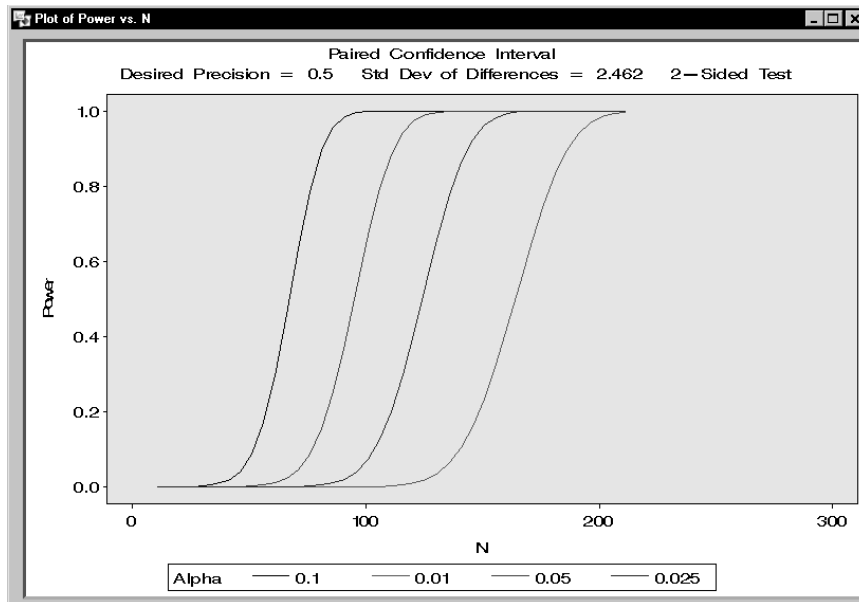
Figure 12.7 contains the completed dialog. Note that, because multiple alpha values were entered, sets of results will be created for each one.



Paired Confidence Interval		
Desired Precision = 0.5    Std Dev of Differences = 2.462    2-Sided Test		
Alpha	N	Power
0.025	36	<.01
	41	<.01
	46	<.01
	51	<.01
	56	<.01
	61	<.01
	66	<.01
	71	<.01
	76	<.01
	81	<.01
	86	0.010
	91	0.020
	96	0.040
	101	0.073
	106	0.125
	111	0.199
	116	0.298
	121	0.415
	126	0.543
	131	0.668
136	0.777	
141	0.863	
146	0.924	
151	0.961	
156	0.982	
161	>.99	
166	>.99	
171	>.99	

**Figure 12.8.** Sample Size Results for Paired Confidence Interval

Figure 12.8 contains the power computations for the sample sizes ranging from 36 to 171. The power analysis results in Figure 12.8 show that, for a two-sided paired confidence interval where the standard deviation of the differences is 2.462, the significance level is  $\alpha = 0.025$ , the sample size is 121, and the power is 0.415. That is, 0.415 represents the probability that a confidence interval containing the true parameter value has a length of no more than twice the desired precision of 0.5.



**Figure 12.9.** Plot for Paired Confidence Interval

The requested plot is displayed in [Figure 12.9](#) and includes one power curve for each specified alpha value.

## Equivalence Tests

In a test of equivalence, a treatment mean and a reference mean are compared to each other. Equivalence is taken to be the alternative hypothesis, and the null hypothesis is nonequivalence. The model assumed may be additive or multiplicative. In the additive model (Phillips 1990), the focus is on the difference between the treatment mean and the reference mean, while in the multiplicative model (Diletti, Hauschke, and Steinijans 1991), the focus is on the ratio of the treatment mean to the reference mean.

In the additive model, the null hypothesis is that the difference between the treatment mean and the reference mean is not near zero. That is, the dif-

ference is less than the lower equivalence bound or greater than the upper equivalence bound and thus nonequivalent.

The alternative is that the difference is between the equivalence bounds; therefore, the two means are considered to be equivalent.

In the multiplicative model, the null hypothesis is that the ratio of the treatment mean to the reference mean is not near one. That is, the ratio is below the lower equivalence bound or above the upper equivalence bound, and thus the two means are not equivalent. The alternative is that the ratio is between the bounds; thus, the two means are considered to be equivalent.

The power of a test is the probability of rejecting the null hypothesis when the alternative is true. In this case, the power is the probability of accepting equivalence when the treatments are in fact equivalent, that is, the treatment difference or ratio is within the prespecified boundaries.

Often, the null difference is specified to be 0; the null hypothesis is that the treatment difference is less than the lower bound or greater than the upper bound, and the alternative is that the difference is not outside the bounds specified. However, in a case where you suspect that the treatments differ slightly (for example,  $\mu_1 = 6$ ,  $\mu_2 = 5$ ,  $\mu_1 - \mu_2 = 1$ ), but you want to rule out a larger difference (for example,  $|\mu_1 - \mu_2| > 2$ ) with probability equal to the power you select, you would specify the null difference to be 1 and the lower and upper bounds to be  $-2$  and  $2$ , respectively. Note that the null difference must lie within the bounds you specify.

### **Requesting Sample Sizes for One Sample In Equivalence**

As an example of computing sample sizes for an equivalence test, consider determining sample sizes for an additive model. The coefficient of variation is 0.2, and the null differences of interest are 0, 0.05, 0.10, and 0.15. The significance level under investigation is 0.05, and the power of interest is 0.80. The lower and upper equivalence bounds are  $-0.2$  and  $0.2$ , respectively.

To perform this computation, select

**Statistics → Sample Size → One-Sample Equivalence . . .**

**Figure 12.10.** Sample Size Dialog for One-Sample Equivalence

Figure 12.10 displays the resulting dialog. For this analysis, you need to input the model type, null difference, coefficient of variation, and the usual alpha level. In addition, you need to specify the equivalence bounds.

These bounds should be chosen to be the minimum difference so that, if the treatments differed by at least this amount, you would consider them to be different. For the multiplicative model, enter the bioequivalence lower and upper limits. For the additive model, enter the bioequivalence lower and upper limits as percentages of the reference mean  $\frac{\text{lowerbound}}{\mu_R}$  and  $\frac{\text{upperbound}}{\mu_R}$ .

For the null difference or ratio, specify one or more values for the null hypothesis difference between the treatment and reference means (additive model) or the ratio of means (multiplicative model). The null difference/ratio value must lie within the equivalence bounds you specify. For the additive model, specify the null difference as a percentage of the reference mean  $\frac{|\mu_T - \mu_R|}{\mu_R}$ , where  $\mu_T$  is the hypothesized treatment mean, and  $\mu_R$  is the hypothesized reference mean. For the multiplicative model, calculate the null ratio as  $\frac{\mu_T}{\mu_R}$ .

You must also input one or more values for the coefficient of variation (c.v.). For the additive model, enter this as a percentage of the reference mean  $\frac{\sigma}{\mu_R}$ ,

which can be estimated by  $\frac{\sqrt{MSE}}{\mu_R}$ . For the multiplicative model, the coefficient of variation is defined as  $\sqrt{e^{(\sigma^2)} - 1}$ . You can estimate  $\sigma$  by  $\hat{\sigma}$ , where  $\hat{\sigma}^2$  is the residual variance of the logarithmically transformed observations. That is,  $\sigma$  can be estimated by  $\sqrt{MSE}$  from the ANOVA of the transformed observations.

To produce sample size computations for the preceding problem, follow these steps:

1. Select **N**.
2. Select **Additive**.
3. Enter 0, 0.05, 0.10, and 0.15 as values for **Null difference**:
4. Enter 0.20 for **Coeff of variation**:
5. Enter 0.05 as the **Alpha**:
6. Enter 0.80 as the **Power**:
7. Enter  $-0.2$  and  $0.2$  as the values for **Lower**: and **Upper**:, respectively, for the **Equivalence bounds**.
8. Click **OK** to perform the analysis.

Figure 12.11 displays the completed dialog.

One-Sample Equivalence

Calculate:  Power  N

Model:  Additive  Multiplicative

Test specifications

Null difference: 0 0.05 0.10 0.15

Coeff of variation: 0.20

Alpha: 0.05

Power: From: 0.80 To: By:

Plot

Power vs. N

Power ref line:

N ref line:

Equivalence bounds

Lower: -0.2

Upper: 0.2

OK

Cancel

Reset

Save Options

Help

Titles

**Figure 12.11.** Sample Size Dialog for One-Sample Equivalence

The results are displayed in [Figure 12.12](#).

Sample Size Analysis

One-Sample Equivalence

Additive Model Lower Bound = -0.2 Upper Bound = 0.2

Coefficient of Variation = 0.20 Alpha = 0.05

Null Difference	Power	N
0.00	0.800	11
0.05	0.800	13
0.10	0.800	27
0.15	0.800	101

**Figure 12.12.** Results for One-Sample Equivalence

The results consist of the sample sizes for a power of 0.80 for the values of the null difference, as displayed in [Figure 12.12](#). These results are for the alpha level of 0.05. For a null difference of 0.10, the sample size is 27. For a null difference of 0.15, the sample size jumps to 101.

---

## One-Way ANOVA

When you are planning to analyze data from more than two groups with a one-way ANOVA, you need to calculate your sample size and power accordingly. These computations are available, prospectively, for use in the planning stages of the study, using the Sample Size task. Retrospective calculations are available, for use in the analysis stage, from the One-Way ANOVA task. This section discusses the prospective computations available in the Analyst Application, which assume equally sized groups.

You must supply two quantities in order to produce these computations: the corrected sum of squares of means (CSS) and the standard deviation. CSS is calculated as

$$CSS = \sum_{g=1}^G (\mu_g - \mu_{.})^2$$

where

$\mu_g$  = mean of the  $g$ th group

$\mu_{.}$  = overall mean

You must enter one or more values for the standard deviation, which in this case is the square root of the Mean Squared Error (MSE).

### **Requesting Power Computations for ANOVA**

The following is an example of calculating the power for a one-way ANOVA with specified values of sample size. Suppose that you are comparing three

groups, the overall mean is 5.5, and the group means are 4.5, 5.5, and 6.5. Therefore, the corrected sum of squares of means (CSS) is

$$(4.5 - 5.5)^2 + (5.5 - 5.5)^2 + (6.5 - 5.5)^2 = 2$$

The standard deviation is the square root of the MSE, which is 1.4142. You are interested in studying sample sizes that range from 6 to 20.

To perform these computations, select

**Statistics → Sample Size → One-Way ANOVA . . .**

Figure 12.13 displays the resulting dialog. For this analysis, you need to enter the number of treatments, or factor levels, the CSS of means, the standard deviation, and the alpha level.

The dialog box is titled "One-Way ANOVA". It contains the following elements:

- Calculate:** Two radio buttons: "Power" (selected) and "N per group".
- Test specifications:**
  - # of treatments: [ ]
  - CSS of means: [ ]
  - Standard deviation: [ ]
  - Alpha: [ 0.05 ]
  - N per group: [ ] [ ] [ ] (labeled From, To, By)
- Plot:**
  - Power vs. N per group
  - Power ref line: [ ]
  - N ref line: [ ]
- Buttons:** OK, Cancel, Reset, Save Options, Help, Titles.

**Figure 12.13.** Sample Size Dialog for One-Way ANOVA



To produce power computations for the preceding problem, follow these steps:

1. Select **Power**.
2. Enter 3 for **# of treatments**:
3. Enter 2 for **CSS of means**:
4. Enter 1.4142 for **Standard deviation**:
5. Enter 0.05 for **Alpha**:
6. Enter 6, 20, and 1 for the fields **N per group**:, for **From**:, **To**:, and **By**:, respectively.
7. Click **OK** to perform the analysis.

Figure 12.14 displays the completed dialog.

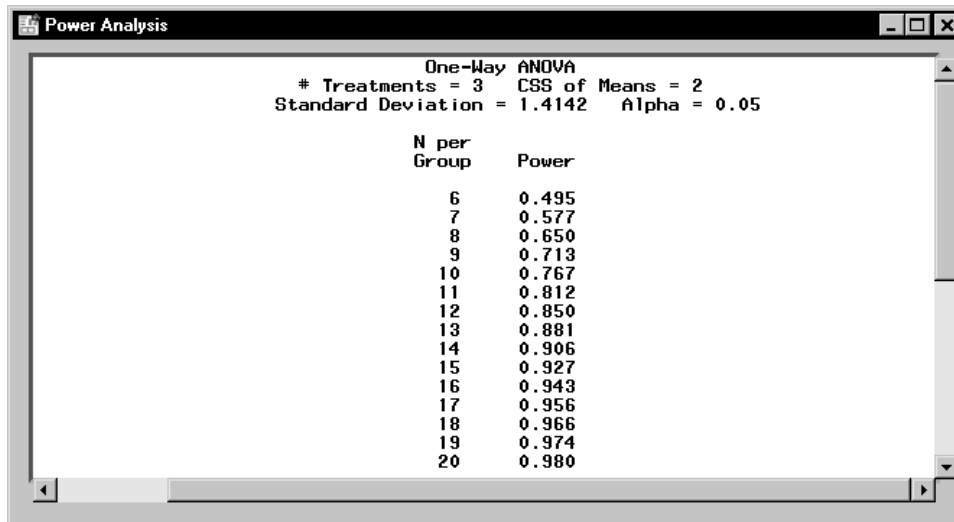
The screenshot shows the 'One-Way ANOVA' dialog box. At the top, there are two radio buttons under 'Calculate:': 'Power' (selected) and 'N per group'. Below this is the 'Test specifications' section with the following fields and values:

# of treatments:	3		
CSS of means:	2		
Standard deviation:	1.4142		
Alpha:	0.05		
N per group:	From:	To:	By:
	6	20	1

At the bottom is the 'Plot' section with a checked box for 'Power vs. N per group' and two empty input fields for 'Power ref line:' and 'N ref line:'. On the right side of the dialog, there are buttons for 'OK', 'Cancel', 'Reset', 'Save Options', 'Help', and 'Titles'.

**Figure 12.14.** Sample Size Dialog for One-Way ANOVA

Requested are power computations for sample sizes ranging from 6 to 20.



**Figure 12.15.** Results for Power Computations for One-Way ANOVA

The results are displayed in Figure 12.15. Note that, to achieve a minimum of 80% power, 11 units per group would be needed.

---

## Power Computation Details

This section provides information on how the power is computed in the Analyst Application. When you request that sample size be computed, the computations produce the smallest sample size that provides the specified power.

---

## Hypothesis Tests

The power for the one-sample *t*-test, the paired *t*-test, and the two-sample *t*-test is computed in the usual fashion. That is, power is the probability of correctly rejecting the null hypothesis when the alternative is true. The sample size is the number per group; these calculations assume equally sized

groups. To compute the power of a  $t$ -test, you make use of the noncentral  $t$  distribution. The formula (O'Brien and Lohr 1984) is given by

$$\text{Power} = \text{Prob}(t > t_{crit}, \nu, NC)$$

for a one-sided alternative hypothesis and

$$\text{Power} = \text{Prob}(t > t_{critu}, \nu, NC) + \text{Prob}(t < t_{critl}, \nu, NC)$$

for a two-sided alternative hypothesis where  $t$  is distributed as noncentral  $t(NC, \nu)$ .

$t_{crit} = t_{(1-\alpha, \nu)}$  is the  $(1 - \alpha)$  quantile of the  $t$  distribution with  $\nu$  df

$t_{critu} = t_{(1-\alpha/2, \nu)}$  is the  $(1-\alpha/2)$  quantile of the  $t$  distribution with  $\nu$  df

$t_{critl} = t_{(\alpha/2, \nu)}$  is the  $(\alpha/2)$  quantile of the  $t$  distribution with  $\nu$  df

For one sample and paired samples,

$\nu = n - 1$  is the df

$NC = \delta\sqrt{n}$  is the noncentrality parameter

For two samples,

$\nu = 2(n - 1)$  is the df

$NC = \frac{\delta}{\sqrt{2/n}}$  is the noncentrality parameter

Note that  $n$  equals the sample size (number per group).

The other parameters are

$$\delta = \begin{cases} \frac{|\mu_a - \mu_0|}{s} & \text{for one-sample} \\ \frac{(\mu_1 - \mu_2)}{s} & \text{for two-sample and paired samples} \end{cases}$$

$$s = \begin{cases} \text{standard deviation for one-sample} \\ \text{standard deviation of the differences for paired samples} \\ \text{pooled standard deviation for two samples} \end{cases}$$

---

## One-Way ANOVA

The power for the one-way ANOVA is computed in a similar manner as for the hypothesis tests. That is, power is the probability of correctly rejecting the null (all group means are equal) in favor of the alternative hypothesis (at least one group mean is not equal), when the alternative is true. The sample size is the number per group; these calculations assume equally sized groups. To compute the power, you make use of the noncentral  $F$  distribution. The formula (O'Brien and Lohr 1984) is given by

$$\text{Power} = \text{Prob}(F > F_{crit}, \nu_1, \nu_2, NC)$$

where  $F$  is distributed as the noncentral  $F(NC, \nu_1, \nu_2)$  and  $F_{crit} = F_{(1-\alpha, \nu_1, \nu_2)}$  is the  $(1-\alpha)$  quantile of the  $F$  distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom.

$\nu_1 = r - 1$	is the numerator df
$\nu_2 = r(n - 1)$	is the denominator df
$n$	is the number per group
$r$	is the number of groups
$NC = \frac{nCSS}{\sigma^2}$	is the noncentrality parameter
$CSS = \sum_{g=1}^G (\mu_g - \mu.)^2$	is the corrected sum of squares
$\mu_g$	is the mean of the $g$ th group
$\mu.$	is the overall mean
$\sigma^2$	is estimated by the mean squared error (MSE)

---

## Confidence Intervals

Power calculations are available when the proposed analysis is construction of confidence intervals of a mean (one-sample) or difference of two means (two-samples or paired-samples). To understand the power of a confidence interval, first define the *precision* to be half the length of a two-sided confidence interval (or the distance between the endpoint and the parameter estimate in a one-sided interval). The power can then be considered to be the

probability that the desired precision is achieved, that is, the probability that the length of the two-sided interval is no more than twice the desired precision. Here, a slight modification of this concept is used. The power is considered to be the conditional probability that the desired precision is achieved, given that the interval includes the true value of the parameter of interest. The reason for the modification is that there is no reason to want the interval to be particularly small if it does not contain the true value of the parameter.

To compute the power of a confidence interval or an equivalence test, you make use of Owen's Q formula (Owen 1965). The formula is given by

$$Q_\nu(t, \delta; a, b) = \frac{\sqrt{2\pi}}{\Gamma(\frac{\nu}{2})2^{(\nu-2)/2}} \int_a^b \Phi\left(\frac{tx}{\sqrt{\nu}} - \delta\right) x^{\nu-1} \phi(x) dx$$

where

$$\Phi = \int_{-\infty}^x \phi(t) dt$$

and

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{(-x^2/2)}$$

The power of a confidence interval (Beal 1989) is given by

$$\text{Power} = \frac{2[Q_\nu(t_c, 0; 0, B) - Q_\nu(0, 0; 0, B)]}{1 - \alpha_s}$$

where

$t_c = t_{(1-\alpha_s/2), \nu}$  is the  $(1 - \alpha_s/2)$  quantile of a  $t$  distribution with  $\nu$  df  
 $\alpha$  is the confidence level

$$\alpha_s = \begin{cases} \alpha & \text{for a two-sided confidence interval} \\ 2\alpha & \text{for a one-sided confidence interval} \end{cases}$$

$$B = \frac{\delta\sqrt{\nu}}{t_c\kappa}$$

$$\left. \begin{array}{l} \nu = n - 1 \\ \kappa = \sqrt{1/n} \end{array} \right\} \text{for the one-sample and paired confidence intervals}$$

$$\left. \begin{array}{l} \nu = 2(n - 1) \\ \kappa = \sqrt{2/n} \end{array} \right\} \text{for the two-sample confidence interval}$$

$\delta = \frac{\text{desired precision}}{\text{standard deviation}}$  is the upper bound of the interval half-length

---

## Equivalence Tests

In a test of equivalence, a treatment mean and a reference mean are compared to each other. Equivalence is taken to be the alternative hypothesis, and the null hypothesis is nonequivalence. The power of a test is the probability of rejecting the null hypothesis when the alternative is true, so in this case, the power is the probability of failing to reject equivalence when the treatments are in fact equivalent, that is, the treatment difference or ratio is within the prespecified boundaries.

The computational details for the power of an equivalence test (refer to Phillips 1990 for the additive model; Diletti, Hauschke, and Steinijans 1991 for the multiplicative) are as follows:

$$\text{Power} = \text{Prob}(t_1 \geq t_{(1-\alpha, \nu)} \text{ and } t_2 \leq -t_{(1-\alpha, \nu)} | \text{bioequivalence})$$

Owen (1965) showed that  $(t_1, t_2)$  has a bivariate noncentral  $t$  distribution that can be calculated as the difference of two definite integrals (Owen's Q function):

$$\text{Power} = Q_\nu(-t_{(1-\alpha, \nu)}, \delta_2; 0, \mathbf{R}) - Q_\nu(t_{(1-\alpha, \nu)}, \delta_1; 0, \mathbf{R})$$

where  $t_{(1-\alpha, \nu)}$  is the  $(1 - \alpha)$  quantile of a  $t$  distribution with  $\nu$  df.

$$\nu = \begin{cases} n - 1 & \text{for the one-sample and paired tests} \\ 2(n - 1) & \text{for the two-sample test} \end{cases}$$

and

$$\left. \begin{array}{l} \delta_1 = \frac{\theta - b_l}{V \cdot \kappa} \\ \delta_2 = \frac{\theta - b_u}{V \cdot \kappa} \\ \theta = \text{null difference} \end{array} \right\} \text{for the additive model}$$

$$\left. \begin{array}{l} \delta_1 = \frac{\log(\theta) - \log(b_l)}{\kappa \sqrt{\log(1+V^2)}} \\ \delta_2 = \frac{\log(\theta) - \log(b_u)}{\kappa \sqrt{\log(1+V^2)}} \\ \theta = \text{null ratio} \end{array} \right\} \text{for the multiplicative model}$$

$V$  is the coefficient of variation

$b_l$  is the lower equivalence bound

$b_u$  is the upper equivalence bound

$$\kappa = \begin{cases} \sqrt{1/n} & \text{for the one-sample and paired tests} \\ \sqrt{2/n} & \text{for the two-sample test} \end{cases}$$

$$R = \frac{\sqrt{\nu}(\delta_1 - \delta_2)}{2 \cdot t_{(1-\alpha, \nu)}}$$

For equivalence tests, alpha is usually set to 0.05, and power ranges from 0.70 to 0.90 (often set to 0.80).

For the **additive model** of equivalence, the values you must enter for the null difference, the coefficient of variation (c.v.), and the lower and upper bioequivalence limits must be expressed as percentages of the reference mean. More information on specifications follow:

Calculate the null difference as  $\frac{|\mu_T - \mu_R|}{\mu_R}$ , where  $\mu_T$  is the hypothesized treatment mean and  $\mu_R$  is the hypothesized reference mean. The null difference is often in the range of 0 to 0.20.

For the coefficient of variation value,  $\sigma$  can be estimated by  $\hat{\sigma}$ , where  $\hat{\sigma}^2$  is the residual variance of the observations (MSE). Enter the c.v. as a percentage of the reference mean, so for the c.v., enter  $\frac{\hat{\sigma}}{\mu_R}$ , or  $\frac{\sqrt{MSE}}{\mu_R}$ . This value is often in the range of 0.05 to 0.30.

Enter the bioequivalence lower and upper limits as percentages of the reference mean as well. That is, for the bounds, enter  $\frac{\text{lowerbound}}{\mu_R}$  and  $\frac{\text{upperbound}}{\mu_R}$ . These values are often  $-0.2$  and  $0.2$ , respectively.

For the **multiplicative model** of equivalence, calculate the null ratio as  $\frac{\mu_T}{\mu_R}$ , where  $\mu_T$  is the hypothesized treatment mean and  $\mu_R$  is the hypothesized reference mean. This value is often in the range of 0.80 to 1.20. More information on specifications follow:

The coefficient of variation (c.v.) is defined as  $\sqrt{e^{(\sigma^2)} - 1}$ . You can estimate  $\sigma$  by  $\hat{\sigma}$ , where  $\hat{\sigma}^2$  is the residual variance of the logarithmically transformed observations. That is,  $\sigma$  can be estimated by  $\sqrt{MSE}$  from the ANOVA of the transformed observations. The c.v. value is often in the range of 0.05 to 0.30.

The bioequivalence lower and upper limits are often set to 0.80 and 1.25, respectively.

---

## References

- Beal, S.L. (1989), "Sample Size Determination for Confidence Intervals on the Population Mean and on the Differences Between Two Population Means," *Biometrics*, 45, 969–977.
- Diletti, E., Hauschke, D., and Steinijans, V.W. (1991), "Sample Size Determination for Bioequivalence Assessment by Means of Confidence Intervals," *International Journal of Clinical Pharmacology, Therapy and Toxicology*, Vol. 29, 1–8.



- O'Brien, R., and Lohr, V. (1984), "Power Analysis For Linear Models: The Time Has Come," *Proceedings of the Ninth Annual SAS User's Group International Conference*, 840–846.
- Owen, D.B. (1965), "A Special Case of a Bivariate Non-central  $t$ -distribution," *Biometrika*, 52, 437–446.
- Phillips, K.F. (1990), "Power of the Two One-Sided Tests Procedure in Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics*, Vol. 18, No. 2, 137–144.



# Chapter 13

## Multivariate Techniques

### Chapter Contents

---

<b>Introduction</b> . . . . .	357
<b>Principal Components Analysis</b> . . . . .	358
<b>Canonical Correlation</b> . . . . .	367
<b>References</b> . . . . .	376



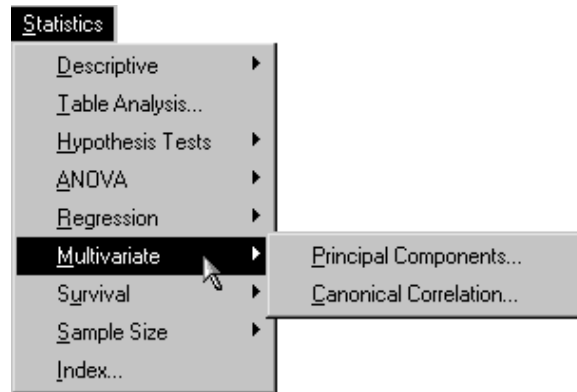
# Chapter 13

## Multivariate Techniques

---

### Introduction

Multivariate analysis techniques, such as principal components analysis and canonical correlation, enable you to investigate relationships in your data. Unlike statistical modeling, you do this without designating dependent or independent variables. In principal component analysis, you examine relationships within a single set of variables. In canonical correlation analysis, you examine the relationship between two sets of variables.



**Figure 13.1.** Multivariate Menu

The Analyst Application enables you to perform principal components analysis and canonical correlation. The Principal Components task enables you to compute principal components from a single set of variables. The Canonical Correlation task enables you to examine the relationship between two sets of variables.

The examples in this chapter demonstrate how you can use the Analyst Application to perform principal components and canonical correlation analyses.

---

## Principal Components Analysis

The purpose of principal component analysis is to derive a small number of independent linear combinations (principal components) of a set of variables that retain as much of the information in the original variables as possible.

For example, suppose you are interested in examining the relationship among measures of food consumption from different sources. The sample data set `Protein` records the amount of protein consumed from nine food groups for each of 25 European countries. The nine food groups are red meat (`RedMt`), white meat (`WhiteMt`), eggs (`Eggs`), milk (`Milk`), fish (`Fish`), cereal (`Cereal`), starch (`Starch`), nuts (`Nuts`), and fruits and vegetables (`FruVeg`).

### *Open the Protein Data Set*

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select `Protein`.
3. Click **OK** to create the sample data set in your `Sasuser` directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select `Sasuser` from the list of **Libraries**.
6. Select `Protein` from the list of members.
7. Click **OK** to bring the `Protein` data set into the data table.

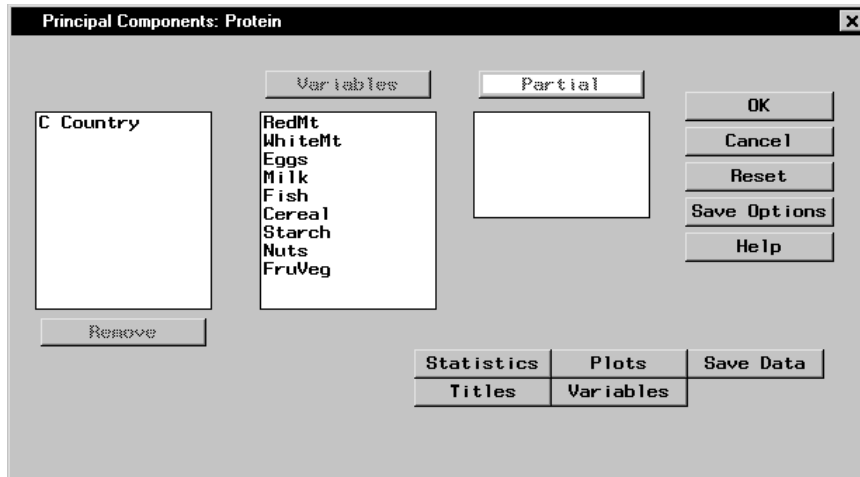
### *Request the Principal Components Analysis*

To perform a principal components analysis, follow these steps:

1. Select **Statistics** → **Multivariate** → **Principal Components** . . .
2. Highlight all of the quantitative variables (`RedMt`, `WhiteMt`, `Eggs`, `Milk`, `Fish`, `Cereal`, `Starch`, `Nuts`, and `FruVeg`).
3. Click on the **Variables** button.

The goal of this analysis is to determine the principal components of all protein sources. Therefore, all of the protein source variables are included in the **Variables** list, as displayed in Figure 13.2. The character variable Country is an identifier variable and is omitted from the **Variables** list.

Note that you can analyze a partial correlation or covariance matrix by specifying the variables to be partialled out in the **Partial** list. The full correlation matrix is used for this analysis.



**Figure 13.2.** Principal Components Dialog

The default principal components analysis includes simple statistics, the correlation matrix for the analysis variables, and the associated eigenvalues and eigenvectors.

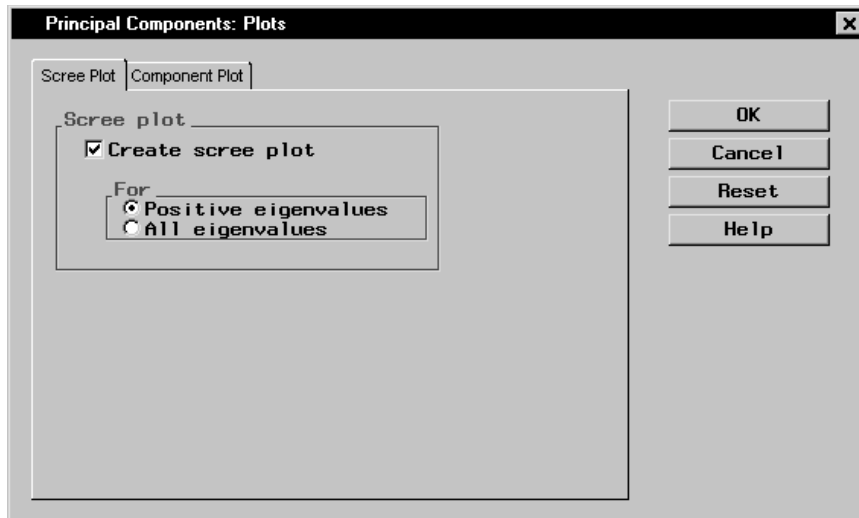
### **Request Principal Component Plots**

You can use the Plots dialog to request a scree plot or component plots. A scree plot is useful in determining the appropriate number of components to interpret. It displays the eigenvalues on the vertical axis and the principal component number on the horizontal axis.

To request a scree plot, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Create scree plot**.

Figure 13.3 displays the **Scree Plot** tab, in which a scree plot of the positive eigenvalues is requested.



**Figure 13.3.** Principal Components: Plots Dialog, Scree Plot Tab

A component plot displays the component score of each observation for a pair of components. When you specify an Id variable, the values of that variable are also displayed in the plot.

To request a component plot in addition to the scree plot, follow these steps.

1. Click on the **Component Plot** tab in the Plots dialog.
2. Select **Create component plots**.
3. Click on the down arrow in the box labeled **Type**:
4. Select **Enhanced**. An enhanced component plot displays the variable names and values of the Id variable in the plot.

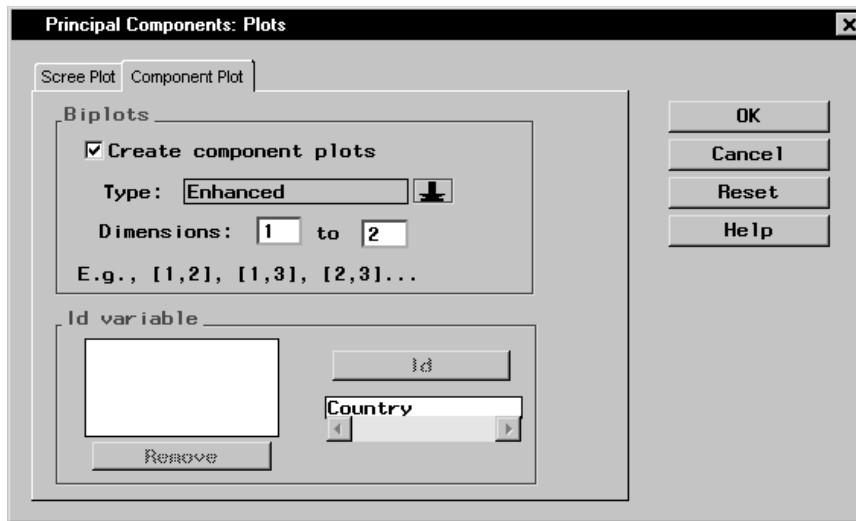


5. Select the variable **Country** in the **Id variable** list.
6. Click on the **Id** button to select the variable **Country** as an Id variable.

You can also enter the **Dimensions** for which you want plots. For example, to request plots of the first versus second, first versus third, and second versus third principal components, you type the values 1 and 3.

7. Click **OK**.

Figure 13.4 displays the **Component Plot** tab, which requests an enhanced component plot.



**Figure 13.4.** Principal Components: Plots Dialog, Component Plot Tab  
Click **OK** in the Principal Components dialog to perform the analysis.

### **Review the Results**

Figure 13.5 displays simple statistics and correlations among the variables.

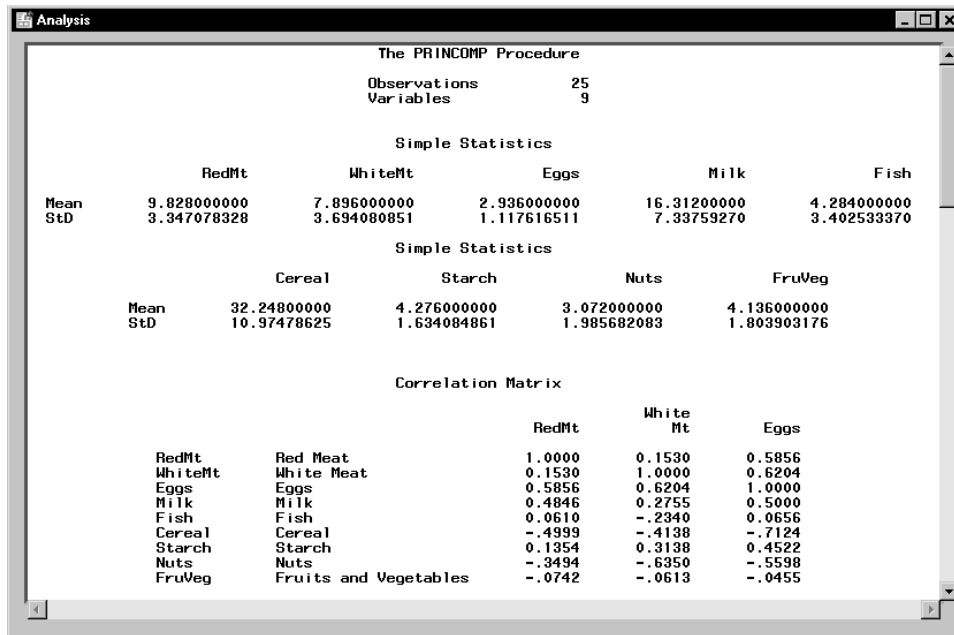
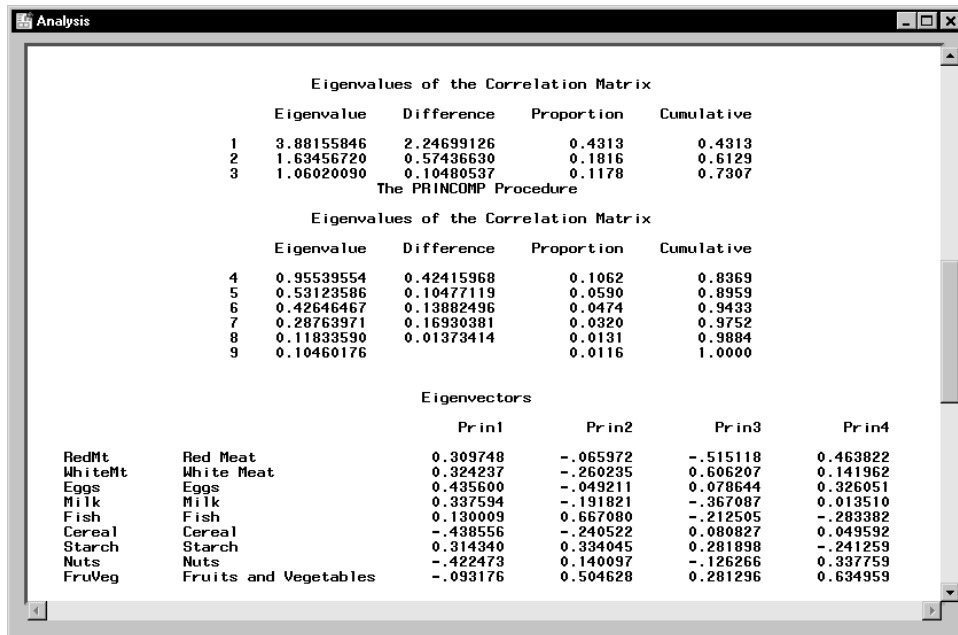


Figure 13.5. Principal Components: Simple Statistics and Correlations

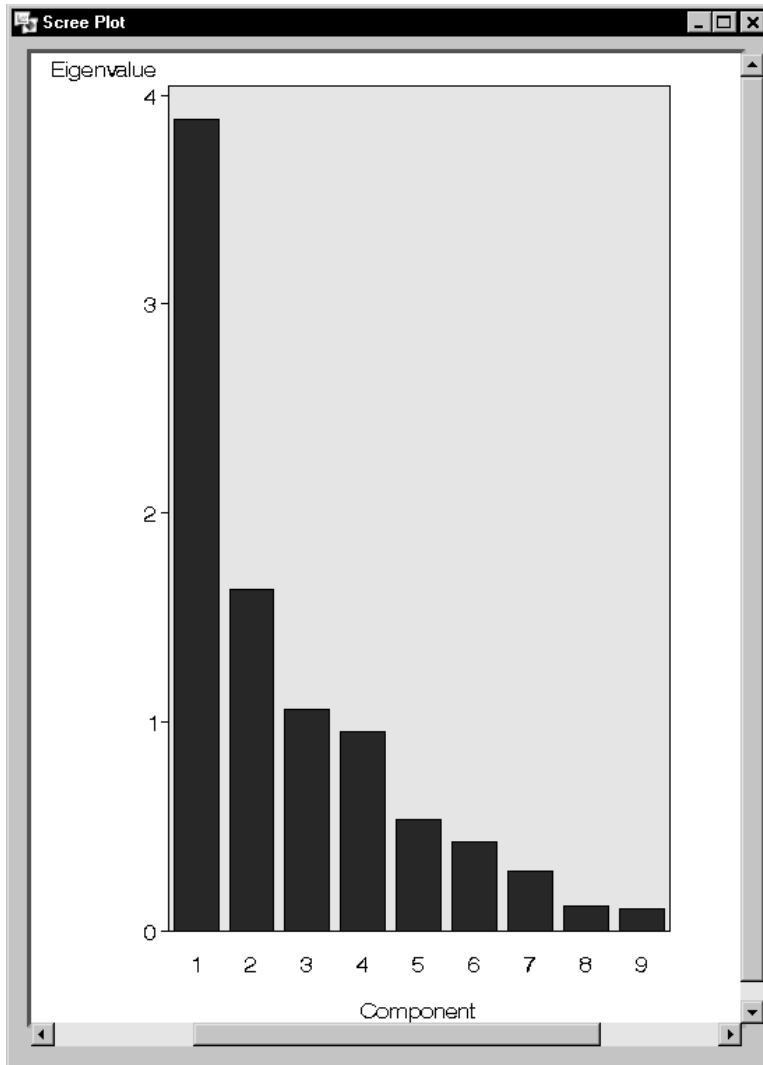


**Figure 13.6.** Principal Components: Eigenvectors and Eigenvalues

Figure 13.6 displays the eigenvalues and eigenvectors of the correlation matrix for the nine variables. The eigenvalues indicate that four components provide a reasonable summary of the data, accounting for about 84% of the total variance. Subsequent components each contribute 5% or less.

The table of eigenvectors in Figure 13.6 reveals that the first eigenvector has equally large loadings on all of the animal-protein variables. This suggests that the first component is primarily a measure of animal-protein consumption. This eigenvector also has a large loading on the variable **Starch** and negative loadings on the variables **Cereal** and **Nuts**.

The second eigenvector has high positive loadings on the variables **Fish**, **Starch**, and **FruVeg**. This component seems to account for diets in coastal regions or warmer climates. The remaining components are not as easily identified.



**Figure 13.7.** Principal Components: Scree Plot

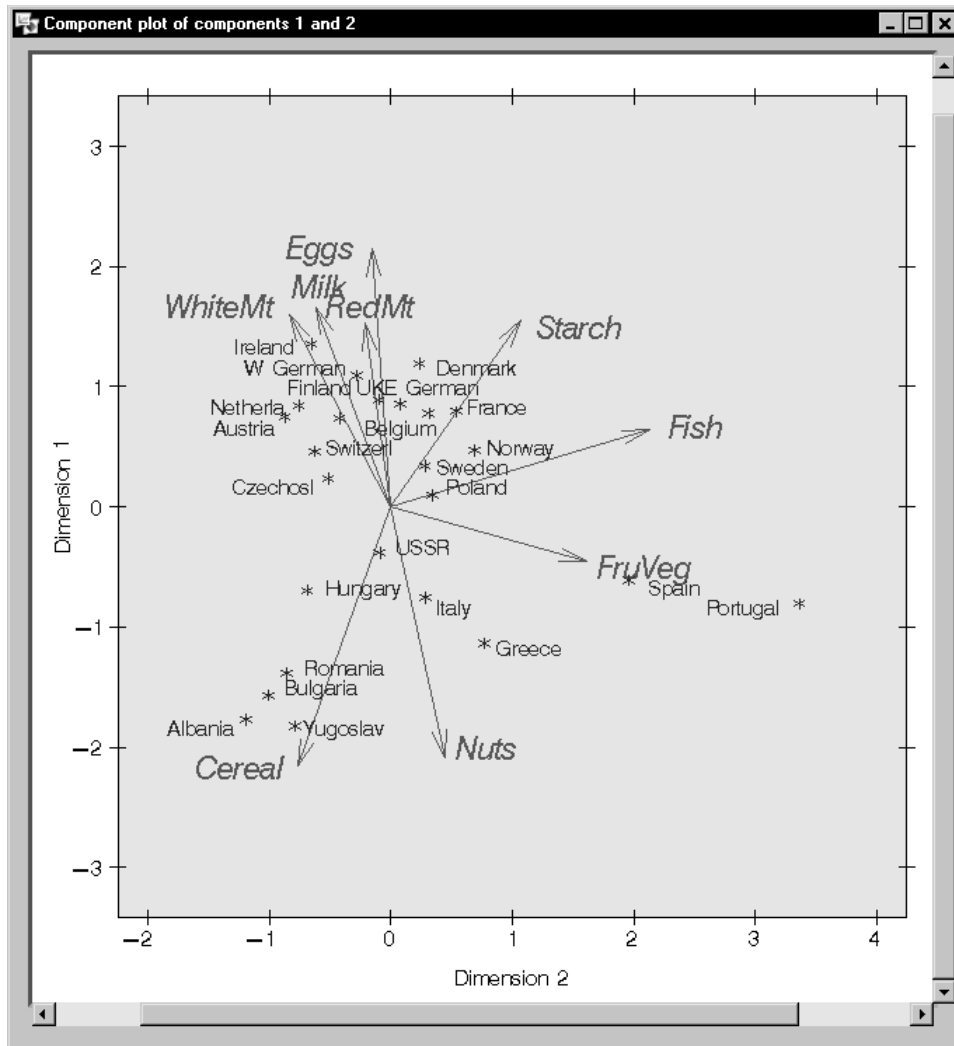
The scree plot displayed in [Figure 13.7](#) shows a gradual decrease in eigenvalues. However, the contributions are relatively low after the fourth component, which agrees with the preceding conclusion that four principal components provide a reasonable summary of the data.

The following enhanced component plot (Figure 13.8) displays the relationship between the first two components; each observation is identified by country.

In addition, the plot is enhanced to depict the correlations between the variables and the components. This correlation is often called the *component loading*. The amount by which each variable “loads” on a component is measured by its correlation with the component.

In Figure 13.8, each vector corresponds to one of the analysis variables and is proportional to its component loading. For example, the variables Eggs, Milk, and RedMt all load heavily on the first component. The variables Fish and FruVeg load heavily on the second component but load very little on the first component.

The information provided by the variable Country reveals that western European countries tend to consume protein from more expensive sources (that is, meat, eggs, and milk), while countries near the Mediterranean Sea rely more heavily on fruits, vegetables, nuts, and fish for their protein sources. Eastern European countries rely more on cereal crops and nuts to supply their protein.



**Figure 13.8.** Principal Components: Scores and Component Loading Plot

---

## Canonical Correlation

Canonical correlation analysis is a variation on the concept of multiple regression and correlation analysis. In multiple regression and correlation analysis, you examine the relationship between a single Y variable and a linear combination of a set of X variables. In canonical correlation analysis, you examine the relationship between a linear combination of the set of Y variables and a linear combination of the set of X variables.

For example, suppose that you want to determine the degree of correspondence between a set of job characteristics and measures of employee satisfaction. The sample data set **Jobs** contains the task characteristics and satisfaction profiles for 14 jobs. The three variables associated with job satisfaction are career track satisfaction (**Career**), management and supervisor satisfaction (**Supervis**), and financial satisfaction (**Finance**). The three variables associated with job characteristics are task variety (**Variety**), supervisor feedback (**Feedback**), and autonomy (**Autonomy**).

In this task, the canonical correlation analysis is performed, labels are specified to identify each set of canonical variables, and a plot of the canonical variables is requested.

### ***Open the Jobs Data Set***

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

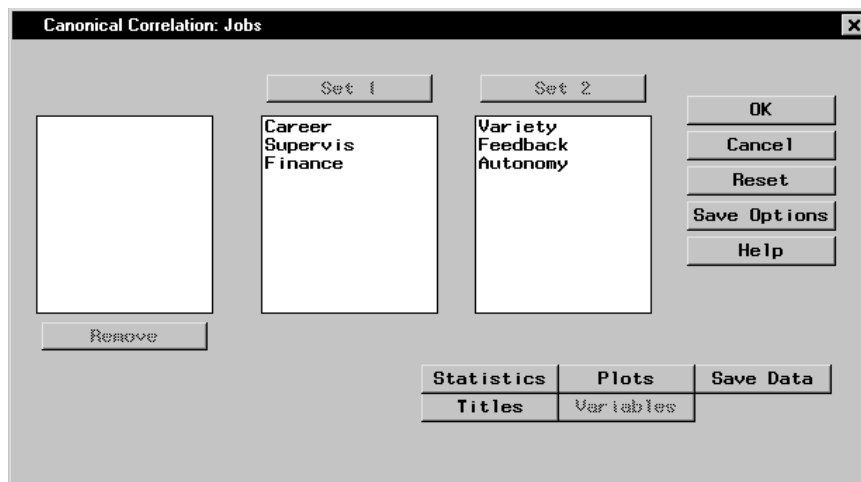
1. Select **Tools** → **Sample Data** . . .
2. Select **Jobs**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Jobs** from the list of members.
7. Click **OK** to bring the **Jobs** data set into the data table.

### Request the Canonical Correlation Analysis

To perform a canonical correlation analysis, follow these steps:

1. Select **Statistics** → **Multivariate** → **Canonical Correlation...**
2. Select the job satisfaction variables (Career, Supervis, and Finance) as the variables in **Set 1**.
3. Select the job characteristic variables (Variety, Feedback, and Autonomy) as the variables in **Set 2**.

Figure 13.9 displays the Canonical Correlation dialog, with each of the two sets of variables defined.



**Figure 13.9.** Canonical Correlation Dialog

The default analysis includes the canonical correlations, eigenvalues, likelihood ratios, and tests of significance.

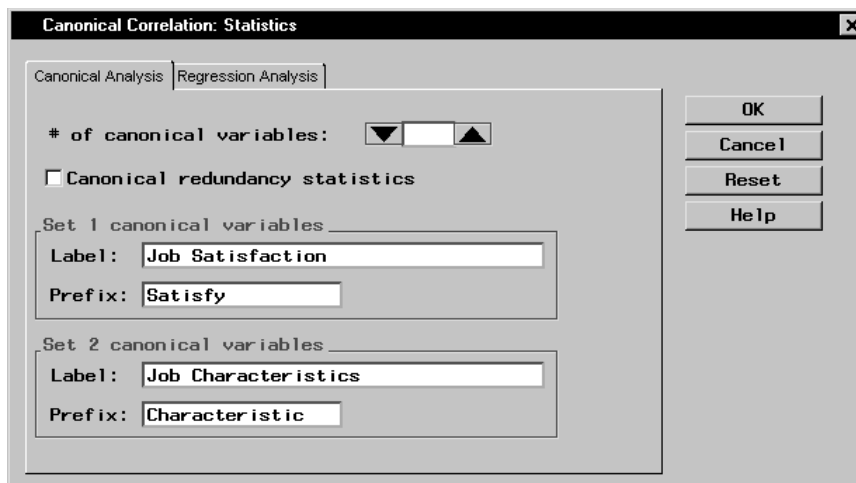
### Specify Identifying Labels

You can optionally specify labels and prefixes to identify the two groups of calculated canonical variables. To specify labels and prefixes, follow these steps:



1. Click on the **Statistics** button in the main dialog.
2. Enter a label for each of the two sets of canonical variables.
3. Enter a prefix for each set of canonical variables. The prefix is used to assign names to the canonical variables.
4. Click **OK**.

Figure 13.10 displays the **Canonical Analysis** tab with labels and prefixes specified.



**Figure 13.10.** Canonical Correlation: Statistics Dialog, Canonical Analysis Tab

### **Request Canonical Variate Plots**

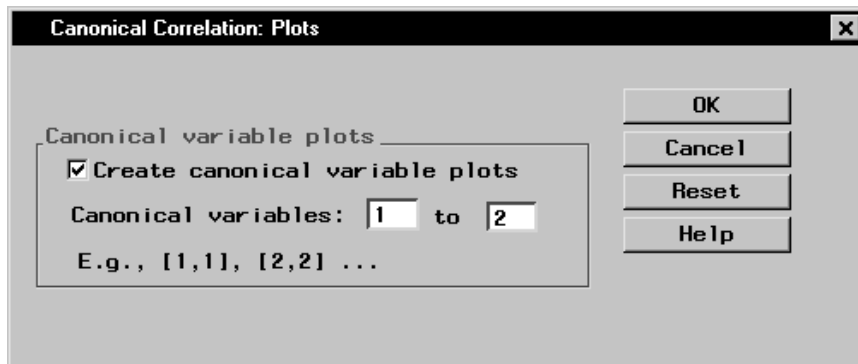
To request plots of the canonical variables, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Create canonical variable plots**.

You can also enter the **Canonical variables** for which you want plots. For example, to request plots of the first, second, and third canonical variable pairs, you would type the values 1 and 3.

3. Click **OK**.

Figure 13.11 displays the Plots dialog, in which plots of the first two canonical variables are requested.

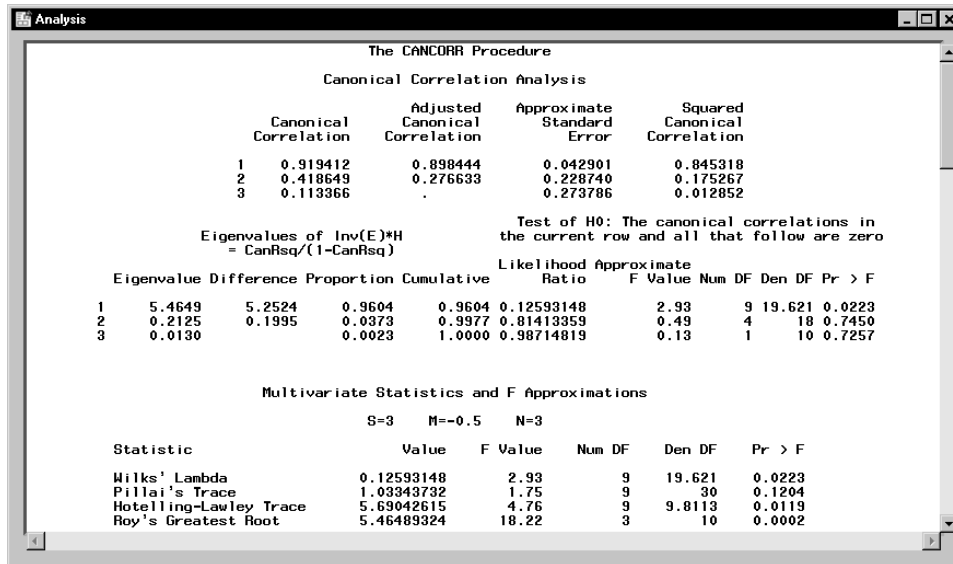


**Figure 13.11.** Canonical Correlation: Plots Dialog

Click **OK** in the Canonical Correlation dialog to perform the analysis.

### **Review the Results**

Figure 13.12 displays the canonical correlation, adjusted canonical correlation, approximate standard error, and squared canonical correlation for each pair of canonical variables.



**Figure 13.12.** Canonical Correlation: Correlations and Eigenvalues

The first canonical correlation (the correlation between the first pair of canonical variables) is 0.9194. This value represents the highest possible correlation between any linear combination of the job satisfaction variables and any linear combination of the job characteristics variables.

Figure 13.12 also displays the likelihood ratios and associated statistics for testing the hypothesis that the canonical correlations in the current row and all that follow are zero. The first approximate  $F$  value of 2.93 corresponds to the test that all three canonical correlations are zero. Since the  $p$ -value is small (0.0223), you can reject the null hypothesis at the  $\alpha = 0.05$  level. The second approximate  $F$  value of 0.49 corresponds to the test that both the second and the third canonical correlations are zero. Since the  $p$ -value is large (0.7450), you fail to reject the hypothesis and conclude that only the first canonical correlation is significant at the  $\alpha = 0.05$  level.

Several multivariate statistics and  $F$  test approximations are also provided. These statistics test the null hypothesis that all canonical correlations are zero. The small  $p$ -values for these tests ( $< 0.05$ ), except for Pillai's Trace, suggest rejecting the null hypothesis that all canonical correlations are zero.

The screenshot shows the following data:

**The CANCORR Procedure**  
**Canonical Correlation Analysis**

**Raw Canonical Coefficients for the Job Satisfaction**

		Satisfy1	Satisfy2	Satisfy3
Career	Career Satisfaction	0.0148378305	-0.026536591	0.0509931964
Supervis	Supervisor Satisfaction	0.0252519157	0.0041970746	-0.02920936
Finance	Financial Satisfaction	0.0243430387	0.4415920204	0.1507204075

**Raw Canonical Coefficients for the Job Characteristics**

		Characteristic1	Characteristic2	Characteristic3
Variety	Task Variety	-0.004300092	0.031408316	0.0351951723
Feedback	Amount of Feedback	0.0201108856	-0.028134366	0.0152825384
Autonomy	Degree of Autonomy	0.0531209636	0.0064473365	-0.052450863

**The CANCORR Procedure**  
**Canonical Correlation Analysis**

**Standardized Canonical Coefficients for the Job Satisfaction**

		Satisfy1	Satisfy2	Satisfy3
Career	Career Satisfaction	0.3028	-0.5416	1.0408
Supervis	Supervisor Satisfaction	0.7854	0.1305	-0.9085
Finance	Financial Satisfaction	0.0538	0.9754	0.3329

**Standardized Canonical Coefficients for the Job Characteristics**

		Characteristic1	Characteristic2	Characteristic3
Variety	Task Variety	-0.1108	0.8095	0.9071
Feedback	Amount of Feedback	0.5520	-0.7722	0.4194
Autonomy	Degree of Autonomy	0.8403	0.1020	-0.8297

**Figure 13.13.** Canonical Correlation: Correlation Coefficients

Even though canonical variables are artificial, they can often be identified in terms of the original variables. To identify the variables, inspect the standardized coefficients of the canonical variables and the correlations between the canonical variables and their original variables. Based on the results displayed in Figure 13.12, only the first canonical correlation is significant. Thus, only the first pair of canonical variables (Satisfy1 and Characteristic1) need to be identified.

The standardized canonical coefficients in Figure 13.13 show that the first canonical variable for the Job Satisfaction group is a weighted sum of the variables Supervis (0.7854) and Career (0.3028), with the emphasis on

Supervis. The coefficient for the variable Finance is near 0. Therefore, a person satisfied with his or her supervisor and with a large degree of career satisfaction would score high on the canonical variable Satisfaction1.

The coefficients for the Job Characteristics variables show that degree of autonomy (Autonomy) and amount of feedback (Feedback) contribute heavily to the Characteristic1 canonical variable (0.8403 and 0.5520, respectively).

Figure 13.14 displays the table of correlations between the canonical variables and the original variables. Although these univariate correlations must be interpreted with caution, since they do not indicate how the original variables contribute jointly to the canonical analysis, they are often useful in the identification of the canonical variables.

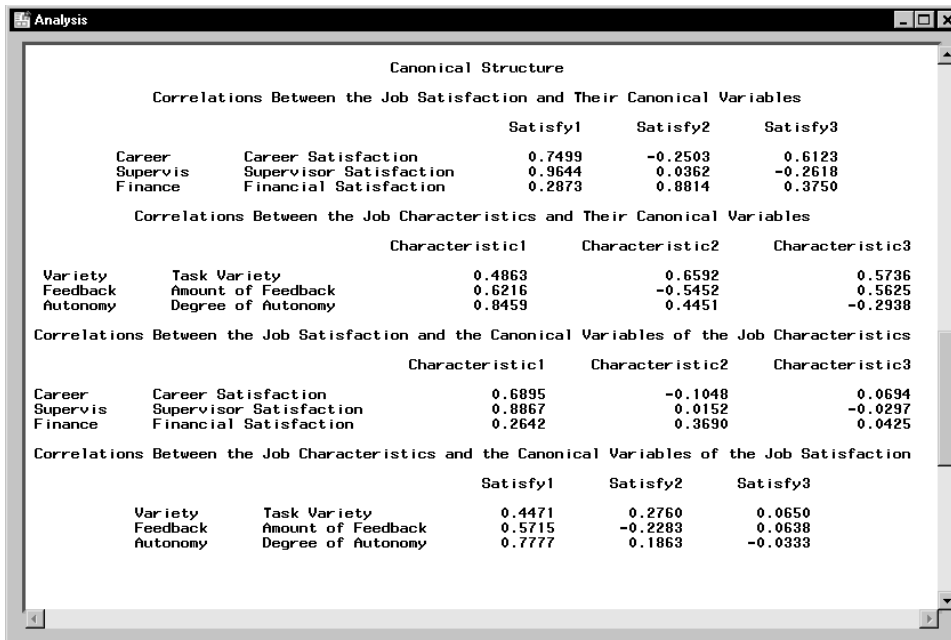


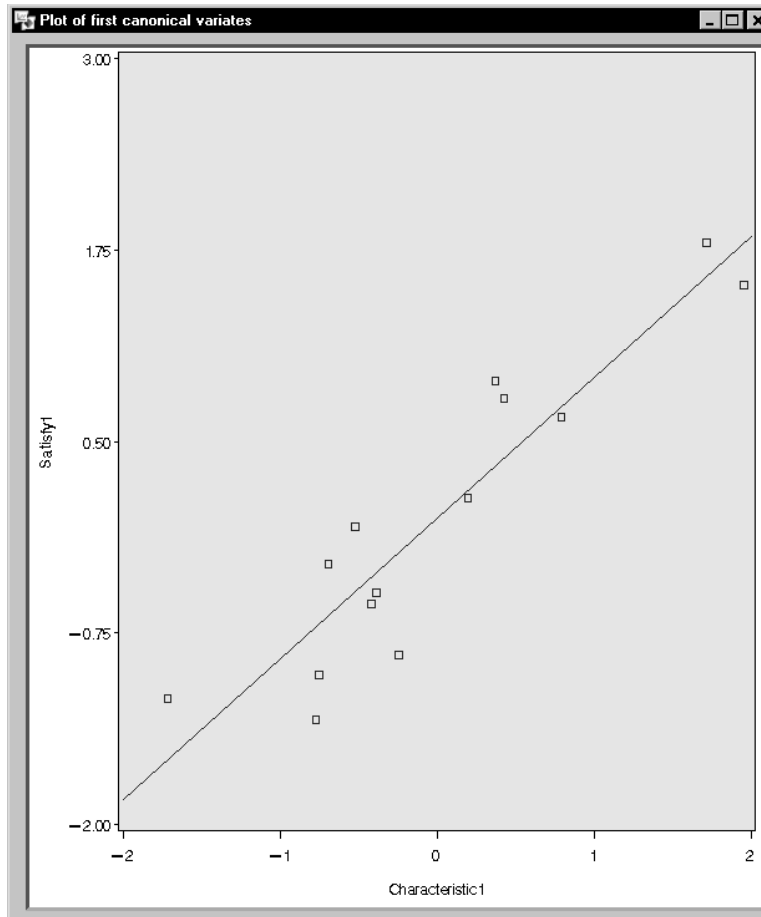
Figure 13.14. Canonical Correlation: Canonical Structure

As displayed in Figure 13.14, the supervisor satisfaction variable, Supervis, is strongly associated with the Satisfy1 canonical variable ( $r = 0.9644$ ).

Slightly less influential is the variable **Career**, which has a correlation with the canonical variable of 0.7499. Thus, the canonical variable **Satisfy1** seems to represent satisfaction with supervisor and career track.

The correlations for the job characteristics variables show that the canonical variable **Characteristic1** seems to represent all three measured variables, with the degree of autonomy variable (**Autonomy**) being the most influential (0.8459).

Hence, you can interpret these results to mean that job characteristics and job satisfaction are related. Jobs that possess a high degree of autonomy and level of feedback are associated with workers who are more satisfied with their supervisors and their careers. Additionally, the analysis suggests that, although the financial component is a factor in job satisfaction, it is not as important as the other satisfaction-related variables.



**Figure 13.15.** Canonical Correlation: Plot of the First Canonical Variables

The plot of the first canonical variables, *Satisfy1* and *Characteristic1*, is displayed in [Figure 13.15](#). The plot depicts the strength of the relationship between the set of job satisfaction variables and the set of job characteristic variables.

---

## References

SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC:  
SAS Institute Inc.



# Chapter 14

## Survival Analysis

### Chapter Contents

---

<b>Introduction</b> . . . . .	379
<b>Life Tables</b> . . . . .	381
<b>Proportional Hazards</b> . . . . .	388
<b>References</b> . . . . .	391



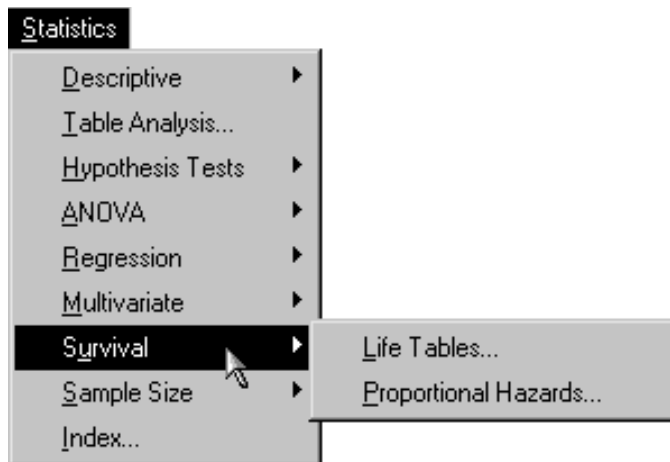
# Chapter 14

## Survival Analysis

---

### Introduction

Survival data often consists of a response variable that measures the duration of time until a specified event occurs and a set of independent variables thought to be associated with the event-time variable. Component lifetimes in industrial reliability, durations of jobs, and survival times in a clinical trial are examples of event times. The purpose of survival analysis is to model the underlying distribution of event times and to assess the dependence of the event time on other explanatory variables. In many situations, the event time is not observed due to withdrawal or termination of the study; this phenomenon is known as *censoring*. Survival analysis methods correctly use both the censored and uncensored observations.



**Figure 14.1.** Survival Analysis Menu

Usually, a first step in the analysis of survival data is the estimation of the distribution of the survival times. The survival distribution function (SDF), also known as the survivor function, is used to describe the lifetimes of the

population of interest. The SDF evaluated at time  $t$  is the probability that an experimental unit from the population will have a lifetime exceeding  $t$ . The product limit and actuarial methods are popular techniques for estimating survival distributions.

Proportional hazards regression is a useful technique for assessing the relationship between survival times and a set of explanatory variables. The proportional hazards model of Cox (1972) assumes a parametric form for the effects of explanatory variables on survival times and allows an unspecified form for the underlying survivor function. The proportional hazards model is also known as Cox regression.

### **Survival Analysis Task Features**

The Life Tables task provides both the actuarial (also known as life-table) method and product-limit method (also known as the Kaplan-Meier method). You can define strata and test the homogeneity of survival functions across strata with rank tests and a likelihood ratio test based on an underlying exponential distribution. In addition, you can test the association between covariates and the lifetime variable with the log-rank test and the Wilcoxon test. Plots provided are the survival function,  $-\log$  (survival function),  $\log(-\log(\text{survival function}))$ , hazard function, and probability density function.

The Proportional Hazards task performs Cox regression. You can choose from five different model selection techniques, select from four different methods for handling tied event times, and produce a survivor function plot with confidence intervals.

The examples in this chapter demonstrate how you can use the Survival tasks in the Analyst Application to analyze survival data.

---

## Life Tables

The data set analyzed in this task contains the survival times of rats in a small randomized trial. Forty rats were exposed to a carcinogen 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 induced by the carcinogen, and interest lies in whether the survival distributions differ between the two treatments. Four rats died of other causes, and their survival times are regarded as censored observations. The data set **Exposed** contains four variables: **Days**, **Status**, **Treatmnt**, and **Gender**. The **Days** variable contains survival times in days from randomization to death, and the **Status** variable has the value 0 for censored observations and 1 for uncensored observations. The **Treatmnt** variable has the value 1 if the rat was administered the first treatment or 2 if the rat was administered the second treatment, and the **Gender** variable has the value F if the rat is female and M if the rat is male.

### *Open the Exposed Data Set*

The data are provided in the Analyst Sample Library. To open the **Exposed** data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select **Exposed**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Exposed** from the list of members.
7. Click **OK** to bring the **Exposed** data set into the data table.

### *Request the Life Tables Analysis*

To specify the Life Tables task, follow these steps.

1. Select **Statistics** → **Survival** → **Life Tables . . .**
2. Select **Days** as the time variable.

A common feature of lifetime or survival data is the presence of right-censored observations due either to withdrawal of experimental units or to termination of the experiment. The analysis methodology must correctly use the censored observations as well as the noncensored observations. In this analysis, the values of **Days** are considered censored if the value of **Status** is 0; otherwise, they are considered event times.

3. Select **Status** as the censoring variable.
4. Specify **0** as the censoring value by directly typing **0** in the **Censoring values:** field or by clicking on the down arrow under **Censoring values:** and selecting **0** from the list. You can remove censoring values by deleting the values in the field.
5. Select **Treatmnt** as the strata variable.

Figure 14.2 displays the dialog with **Days** specified as the time variable, **Status** specified as the censoring variable, **0** selected as the censoring value, and **Treatmnt** specified as the strata variable.

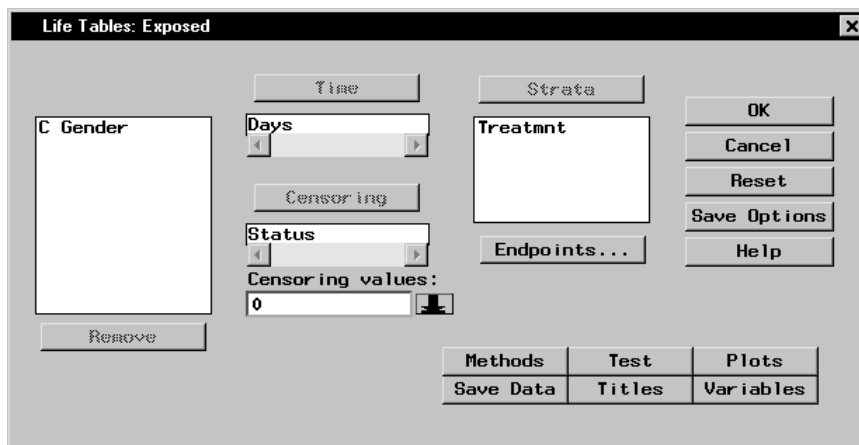
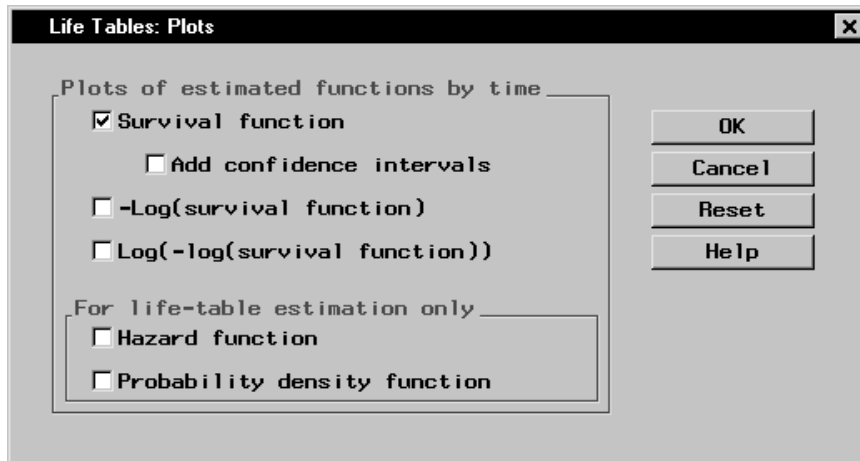


Figure 14.2. Life Tables Dialog

### Request A Survivor Function Plot

To produce a plot of the survivor function, follow these steps:

1. Click **Plots** to open the Plots dialog.
2. Select **Survival function**.
3. Click **OK**.

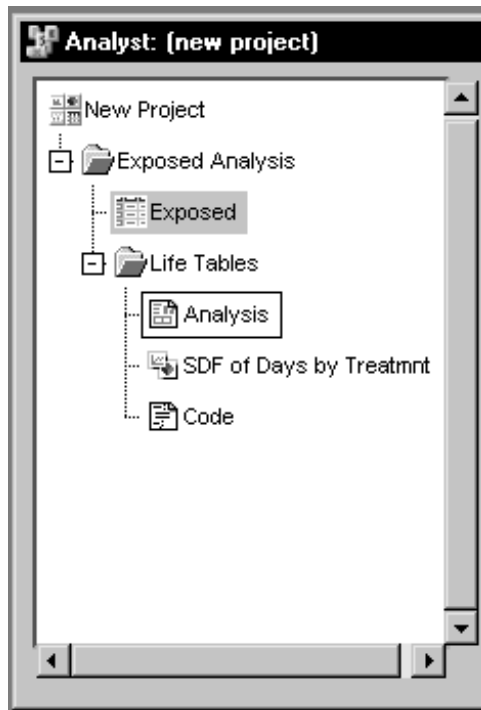


**Figure 14.3.** Life Tables: Plots Dialog

When you have completed your selections, click **OK** in the main dialog to produce the analysis.

### Review the Results

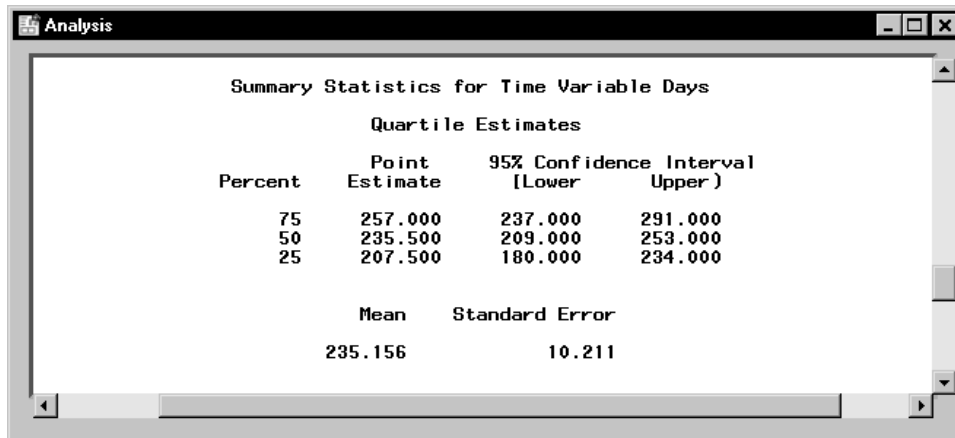
The results are presented in the project tree under the **Life Tables** folder, as displayed in [Figure 14.4](#). The three nodes represent the life tables output, the survivor distribution function plot, and the SAS programming statements (labeled **Code**) that generated the output.



**Figure 14.4.** Life Tables: Project Tree

You can double-click on any node in the project tree to view the contents in a separate window.





The screenshot shows a window titled 'Analysis' with a scrollable area containing the following text:

Summary Statistics for Time Variable Days

Quartile Estimates

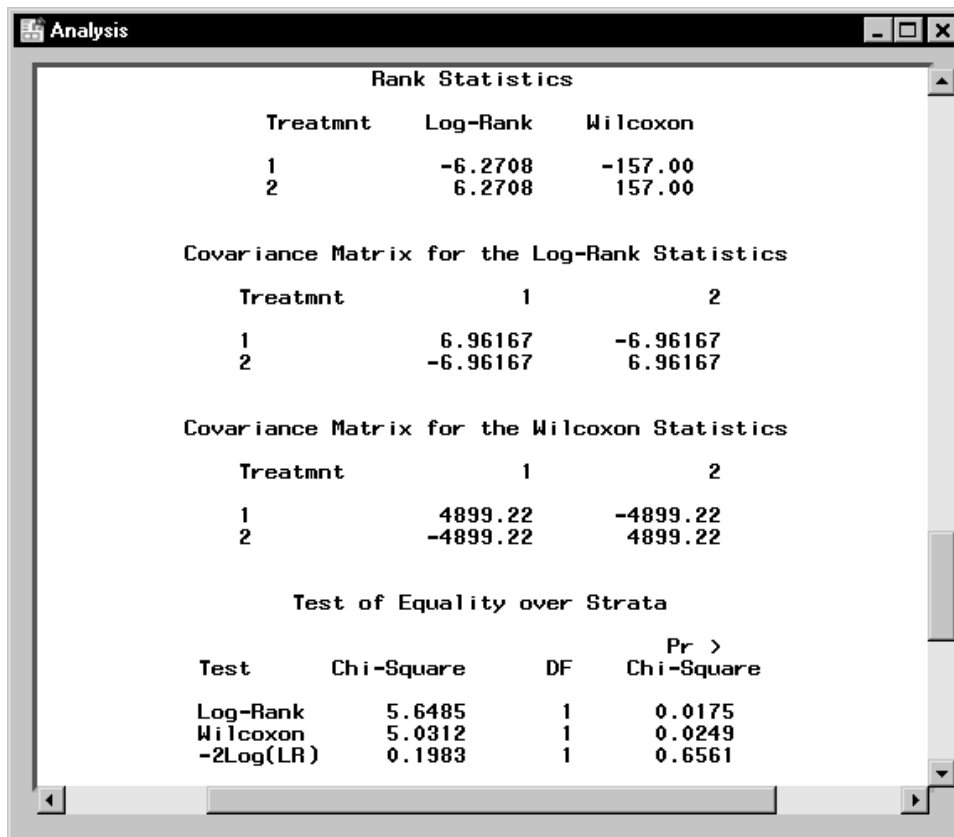
Percent	Point Estimate	95% Confidence Interval (Lower Upper)	
75	257.000	237.000	291.000
50	235.500	209.000	253.000
25	207.500	180.000	234.000

Mean Standard Error

235.156	10.211
---------	--------

**Figure 14.5.** Life Tables: Results

Figure 14.5 displays summary statistics for the survival times for rats administered treatment 2. Of greatest interest is the 50th percentile, which is the median survival time. Here, rats administered treatment 2 have a median survival time of 235.5 days with a 95-percent confidence interval of 209 to 253. The mean survival time is 235.156 with a standard error of 10.211.



**Figure 14.6.** Life Tables: Test for Equality over Strata

The “Test for Equality over Strata” table contains rank and likelihood-based statistics for testing homogeneity of survivor functions across strata. The rank tests for homogeneity indicate a significant difference between the treatments ( $p=0.0175$  for the log-rank test and  $p=0.0249$  for the Wilcoxon test), where rats in the first treatment group live significantly longer than those in the second treatment group. The log-rank test, which places more weight on larger survival times, has a lower  $p$ -value than the Wilcoxon test, which places more weight on early survival times.

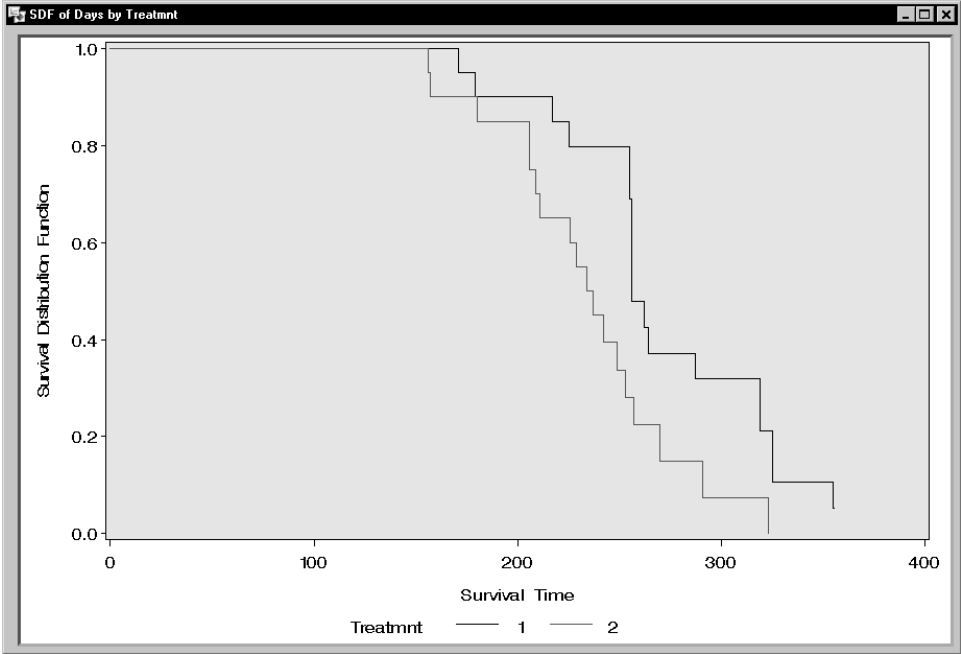


Figure 14.7. Life Tables: Survivor Distribution Plot

Figure 14.7 displays the survivor function against time for each of the two treatments. The gap between the two curves distinguishes between the survival distributions, where the curve for treatment 1 decreases after the curve for treatment 2. The difference in displayed survival curves reinforces the conclusions that the rats in the first treatment group live longer than rats in the second group.

---

## Proportional Hazards

The example in this section contains information on a different study that explores survival times of rats exposed to a carcinogen. Two groups of rats received different pretreatment regimes and were exposed to a carcinogen. Investigators recorded the survival times of the rats from exposure to death from vaginal cancer. Interest lies in whether the survival curves differ between the two groups. The data set **Rats** contains the variables **Days**, **Status**, and **Group**. The variable **Days** is the survival time in days. **Status** is the censoring variable and has the value 0 if the observation is censored and 1 if the observation is not censored. The **Group** variable indicates the pretreatment group, which takes the value 0 for the first treatment and 1 for the second treatment.

### Open the Rats Data Set

The data are provided in the Analyst Sample Library. To open the **Rats** data set, follow these steps:

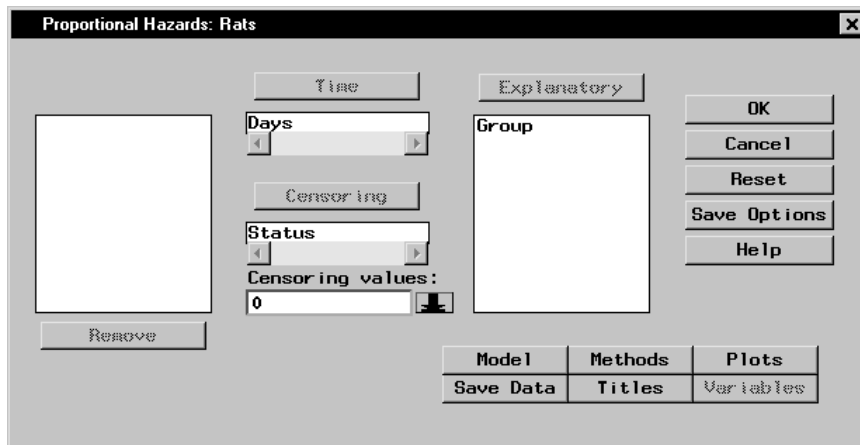
1. Select **Tools** → **Sample Data** . . .
2. Select **Rats**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Rats** from the list of members.
7. Click **OK** to bring the **Rats** data set into the data table.

To request proportional hazards regression, follow these steps:

1. Select **Statistics** → **Survival** → **Proportional Hazards** . . .
2. Select **Days** as the time variable.

The values of **Days** are considered censored if the value of **Status** is 0; otherwise, they are considered event times.

3. Select **Status** as the censoring variable.
4. Specify **0** as the censoring value by directly typing **0** in the **Censoring values:** field or by clicking the down arrow under **Censoring values:** and selecting **0** from the list.
5. Select **Group** as the explanatory variable.



**Figure 14.8.** Proportional Hazards Dialog

Click **OK** in the **Proportional Hazards** main dialog to produce the results for the proportional hazards task.

### **Review the Results**

The results are presented in the project tree under the **Proportional Hazards** folder. Double-click on the icon labeled **Analysis** to display the corresponding information in an independent window.

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	2.8784	1	0.0898	
Score	3.0001	1	0.0833	
Wald	2.9254	1	0.0872	

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Group	1	-0.59590	0.34840	2.9254	0.0872	0.551

**Figure 14.9.** Proportional Hazards: Results

Figure 14.9 displays likelihood statistics and the analysis of parameter estimates. Since **Group** takes only two values, the null hypothesis for no difference between two groups is identical to the null hypothesis that the regression coefficient for **Group** is 0. All three tests in the “Testing Global Null Hypothesis: BETA=0” table suggest that the two pretreatment groups may not be the same. In this model, the hazards ratio (or risk ratio) for **Group**, defined as the exponentiation of the regression coefficient for **Group**, is the ratio of hazard functions between the two groups. The estimate is 0.551, implying that the hazard function for group 1 is smaller than the hazard function for group 0. In other words, rats in group 1 lived longer than those in group 0.

In this example, the comparison of two survival curves is put in the form of a proportional hazards model. This approach is essentially the same as the log-rank (Mantel-Haenszel) test. In fact, if there are no ties in the survival times, the likelihood score test in the Cox regression analysis is identical to the log-rank test. The advantage of the Cox regression approach is the ability to adjust for the other variables by including them in the model. For example, including a variable that contains the initial body weights of the rats could expand the present model.

---

## References

- Allison, P. D. (1995), *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute Inc.
- Cox, D. R. (1972), “Regression Models and Life-Tables (with Discussion),” *Journal of the Royal Statistical Society, Series B*, 34, 187–200.
- SAS Institute Inc. (2000), *SAS/STAT User’s Guide, Version 8*, Cary, NC: SAS Institute Inc.





# Chapter 15

## Mixed Models

### Chapter Contents

---

<b>Introduction</b> . . . . .	395
<b>Split Plot Experiment</b> . . . . .	397
<b>Clustered Data</b> . . . . .	405
<b>References</b> . . . . .	410



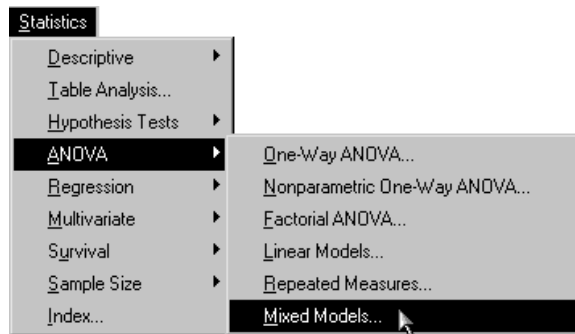
# Chapter 15

## Mixed Models

---

### Introduction

The Mixed Models task provides facilities for fitting a number of basic mixed models. These models enable you to handle both fixed effects and random effects in a linear model for a continuous response. Numerous experimental designs produce data for which mixed models are appropriate, including split-plot experiments, multilocation trials, and hierarchical designs.



**Figure 15.1.** Mixed Models Menu

A standard linear model is designed to handle *fixed effects*, in which the levels of the factor represent all possible levels for that factor or at least all levels about which inference is to be made. Factor effects are *random effects* if the levels of the factor in a study or experiment are randomly selected from a population of possible levels of that factor. The population of possible levels of a random effect has a probability distribution with a mean and a variance. By modeling both fixed and random effects, the mixed model provides you with the flexibility of modeling not only means (as in the standard linear model) but variances and covariances as well.

The mixed model is written

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

where  $\mathbf{y}$  denotes the vector of observed values,  $\mathbf{X}$  is the known fixed effects design matrix, and  $\beta$  is the unknown fixed effects parameter vector.  $\mathbf{Z}\gamma$  represents the additional random component of the mixed model. Here,  $\mathbf{Z}$  is the known random effects design matrix and  $\gamma$  is a vector of unknown random-effects parameters.  $\mathbf{Z}$  contains indicator variables constructed from the random effects, just as  $\mathbf{X}$  contains variables constructed for fixed effects. Finally,  $\epsilon$  is the unobserved vector of independent and identically distributed Gaussian random errors.

Assume that  $\gamma$  and  $\epsilon$  are Gaussian random variables that are uncorrelated and have expectations 0 and variances  $\mathbf{G}$  and  $\mathbf{R}$ , respectively.

$$\begin{aligned} \mathbf{E} \begin{bmatrix} \gamma \\ \epsilon \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \text{Var} \begin{bmatrix} \gamma \\ \epsilon \end{bmatrix} &= \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \end{aligned}$$

The variance of  $\mathbf{y}$  is therefore  $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ .

Note that this is a general specification of the mixed model. The Mixed Models task enables you to specify classification random effects that are a special case of the general specification. You can specify that  $\mathbf{Z}$  contains dummy variables,  $\mathbf{G}$  contains variance components in a diagonal structure, and  $\mathbf{R} = \sigma^2\mathbf{I}_n$ , where  $\mathbf{I}_n$  denotes the  $n \times n$  identity matrix.

The Mixed Models task enables you to specify a mixed model that incorporates fixed effects and random classification effects and includes interactions and nested terms. You can select from six estimation methods, including maximum likelihood, restricted maximum likelihood (REML), and MIVQUE. You can also compute least-squares means, produce Type 1, 2, and 3 tests for fixed effects, and output predicted values and means to a SAS data set. Plots include means plots for fixed effects, predicted plots, and residual plots.

The examples in this chapter demonstrate how you can use the Mixed Models task in the Analyst Application to analyze linear models data that contain fixed and random effects.

---

## Split Plot Experiment

One of the most common mixed models is the split-plot design. The split-plot design involves two experimental factors, A and B. Levels of A are randomly assigned to whole plots (main plots), and levels of B are randomly assigned to split plots (subplots) within each whole plot. The subplots are assumed to be nested within the whole plots so that a whole plot consists of a cluster of subplots and a level of A is applied to the entire cluster. The design provides more precise information about B than about A, and it often arises when A can be applied only to large experimental units.

The hypothetical data set analyzed in this example was created as a balanced split-plot design with the whole plots arranged in a randomized complete-block design (Stroup 1989). The response variable Y represents crop growth measurements. The variable A is a whole plot factor that represents irrigation levels for large plots, and the subplot variable B represents different crop varieties planted in each large plot. The levels of B are randomly assigned to split plots (subplots) within each whole plot. The data set **Split** contains the whole plot factor A, split plot factor B, response Y, and blocking factor **Block**. Using the Mixed Models task, you can estimate variance components for **Block**, **A\*Block**, and the residual and automatically incorporate correct error terms into the tests for fixed effects.

### Open the Split Data Set

These data are provided as the **Split** data set in the Analyst Sample Library. To open the **Split** data set, follow these steps:

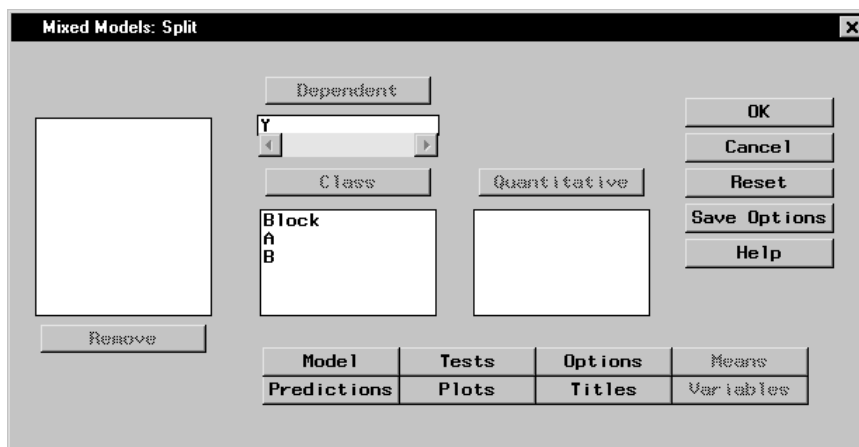
1. Select **Tools** → **Sample Data** . . .
2. Select **Split**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .

5. Select **Sasuser** from the list of **Libraries**.
6. Select **Split** from the list of members.
7. Click **OK** to bring the **Split** data set into the data table.

### Request the Mixed Models Analysis

To specify the split plot analysis, follow these steps:

1. Select **Statistics** → **ANOVA** → **Mixed Models . . .**
2. Select **Y** as the dependent variable.
3. Select **A**, **B**, and **Block** as classification variables.



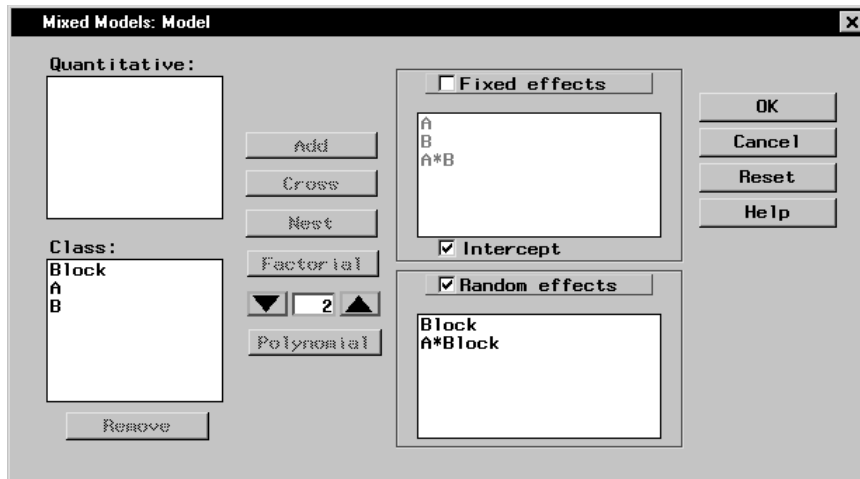
**Figure 15.2.** Mixed Models Dialog

Figure 15.2 displays the dialog with **Y** specified as the dependent variable and **A**, **B**, and **Block** specified as classification effects in the mixed model.

### Specify the Mixed Model

You can define fixed and random effects, create nested terms, and specify interactions in the Model dialog. The Analyst Application adds terms to the **Fixed effects** list or the **Random effects** list depending on whether the

check box at the top of each list is checked. Check the appropriate box for each term you add. Only classification variables can be specified as random effects, and once a term has been specified as a random effect, all higher-order interactions that include that effect must also be specified as random effects.



**Figure 15.3.** Mixed Models: Model Dialog

To specify the mixed model, follow these steps:

1. Click **Model** in the main dialog.
2. Ensure that the **Fixed effects** check box is selected.
3. Select **A** and **B** and click **Factorial**.
4. Select the **Random effects** check box, and then select **Block** and click **Add**.
5. Select **Block** and **A** and click **Cross**.

These selections create a factorial structure that contains the **A** and **B** main effects and the **A\*B** interaction as fixed effects, and **Block** and **A\*Block** as random effects. Since you specified the random effects, the columns of the model matrix **Z** now consist of indicator variables corresponding to the levels

of Block and A\*Block. The **G** matrix is diagonal and contains the variance components of Block and A\*Block; the **R** matrix is also diagonal and contains residual variance.

### Produce Least-Squares Means

You can request generalized least-squares means of fixed effects using the Means dialog. The least-squares means are estimators of the class or subclass marginal means that are expected for a balanced design. Each least-squares mean is computed as  $L\hat{\beta}$ , where **L** is the coefficient matrix associated with the least-squares mean and  $\hat{\beta}$  is the estimate of the fixed-effects parameter vector. Least-squares means can be computed for any fixed effect that is composed of only classification variables.

For this analysis, interest lies in comparing response means across combinations of the levels of A and B. To request least-squares means of the A\*B interaction, follow these steps:

1. Click **Means** in the main dialog.
2. Select A\*B in the candidate list and click **LS Mean**.

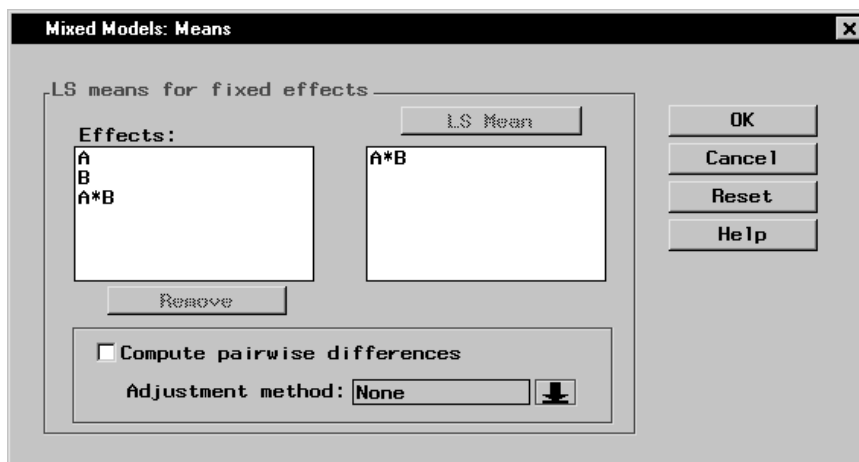


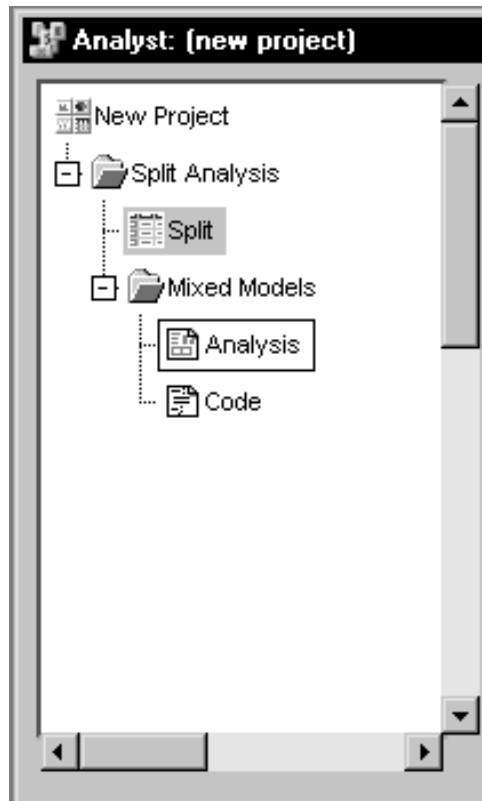
Figure 15.4. Mixed Models: Means Dialog



When you have completed your selections, click **OK** in the main dialog to perform the analysis.

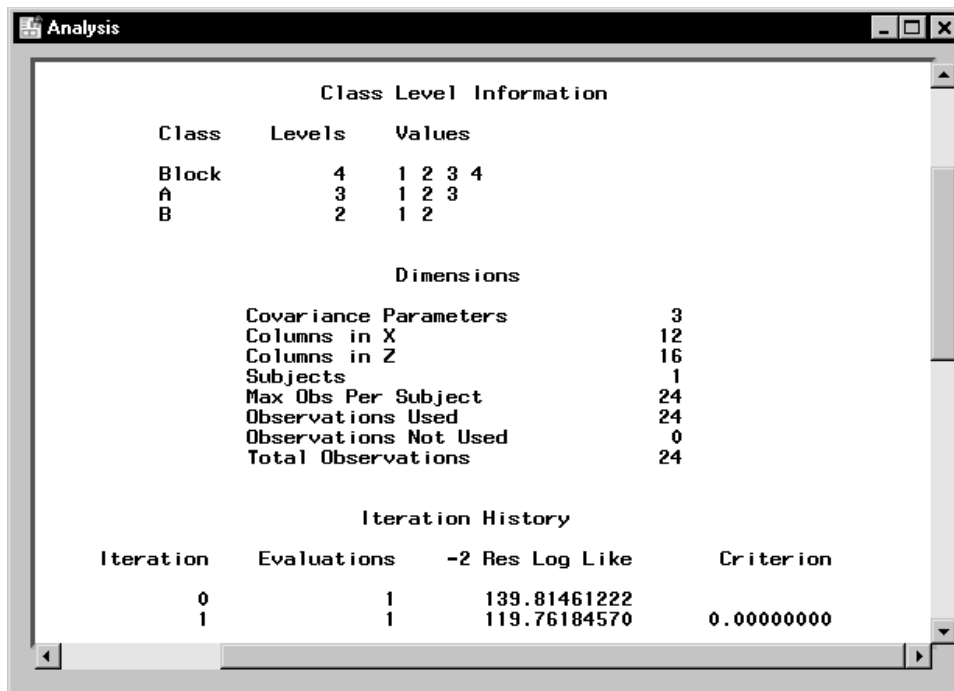
### Review the Results

The results are presented in the project tree under the **Mixed Models** folder, as displayed in [Figure 15.5](#). The two nodes represent the mixed models results and the SAS programming statements (labeled **Code**) that generate the output.



**Figure 15.5.** Mixed Models: Project Tree

Double-click on the **Analysis** node in the project tree to view the contents in a separate window.

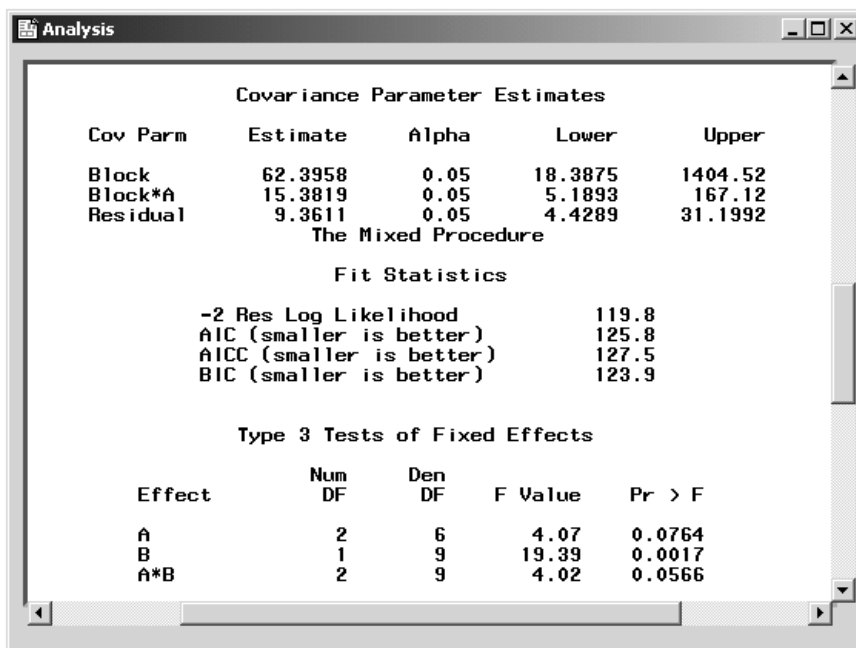


**Figure 15.6.** Mixed Models: Model Information

Figure 15.6 displays class level information, dimensions of model matrices, and the iteration history of the estimated model. The “Class Level Information” table lists the levels of all classification variables included in the model. The “Dimensions” table includes the number of estimated covariance parameters as well as the number of columns in the  $X$  and  $Z$  design matrices.

The Mixed Models task estimates the variance components for **Block**, **A\*Block**, and the residual by a method known as residual (restricted) maximum likelihood (REML). The REML estimates are the values that maximize the likelihood of a set of linearly independent error contrasts, and they provide a correction for the downward bias found in the usual maximum likelihood estimates.

The “Iteration History” table records the steps of the REML optimization process. The objective function of the process is  $-2$  times the restricted likelihood. The Mixed Models task attempts to minimize this objective function via the Newton-Raphson algorithm, which uses the first and second derivatives of the objective function to iteratively find its minimum. For this example, only one iteration is required to obtain the estimates. The Evaluations column reveals that the restricted likelihood is evaluated once for each iteration, and the criterion of 0 indicates that the Newton-Raphson algorithm has converged.

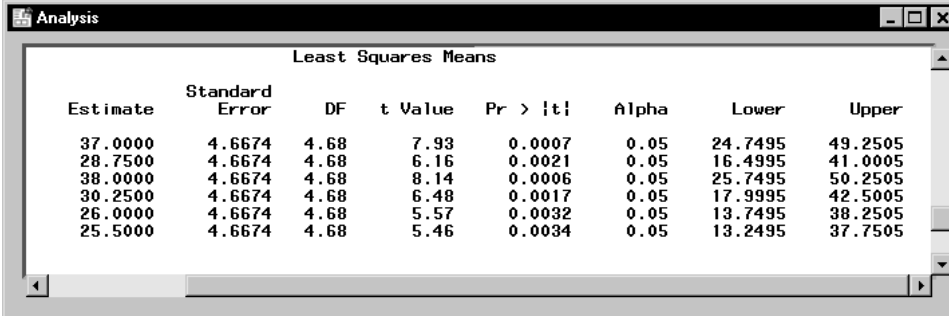


**Figure 15.7.** Mixed Models: Covariance Estimates and Tests for Fixed Effects

Figure 15.7 displays covariance parameter estimates, information on the model fit, and Type 3 tests of fixed effects. The REML estimates for the variance components of Block, A\*Block, and the residual are 62.4, 15.4, and 9.4, respectively. The “Fit Statistics” table lists several pieces of information about the fitted mixed model: the  $-2$  residual log likelihood, Akaike’s

Information Criterion (AIC), a corrected form of AIC that adjusts for small sample size (AICC), and Schwarz's Bayesian Information Criterion (BIC). The information criteria can be used to compare different models; models with smaller values for these criteria are preferred.

The tests of fixed effects are produced using Type 3 estimable functions. The test for the A\*B interaction has a  $p$ -value of 0.0566, indicating that there is moderate evidence of an interaction between crop varieties and irrigation levels.



Estimate	Standard Error	DF	t Value	Pr >  t	Alpha	Lower	Upper
37.0000	4.6674	4.68	7.93	0.0007	0.05	24.7495	49.2505
28.7500	4.6674	4.68	6.16	0.0021	0.05	16.4995	41.0005
38.0000	4.6674	4.68	8.14	0.0006	0.05	25.7495	50.2505
30.2500	4.6674	4.68	6.48	0.0017	0.05	17.9995	42.5005
26.0000	4.6674	4.68	5.57	0.0032	0.05	13.7495	38.2505
25.5000	4.6674	4.68	5.46	0.0034	0.05	13.2495	37.7505

**Figure 15.8.** Mixed Models: Least Squares Means

Figure 15.8 displays the least-squares means for each combination of irrigation levels (A) and crop varieties (B). At each irrigation level, the response is higher for the first crop variety compared to the second variety. The interaction between crop variety and irrigation levels is evident in that variety 1 has a higher mean response than variety 2 at irrigation levels 1 and 2, but the two varieties have nearly the same mean response at irrigation level 3.

---

## Clustered Data

The example in this section contains information on a study investigating the heights of individuals sampled from different families. The response variable **Height** measures the height (in inches) of 18 individuals that are classified according to **Family** and **Gender**. Since the data occurs in clusters (families), it is very likely that observations from the same family are statistically correlated and not independent. In this case, it is inappropriate to analyze the data using a standard linear model.

A simple way to model the correlation is through the use of a **Family** random effect. The **Family** effect is assumed to be normally distributed with mean of zero and some unknown variance. Defining **Family** as a random effect sets up a common correlation among all observations having the same level of family.

In addition, a female within a certain family may exhibit more correlation with other females in that same family than with the males in that family, and likewise for males. Defining **Family\*Gender** as a random effect models an additional correlation for all observations having the same value of both **Family** and **Gender**.

### **Open the Heights Data Set**

These data are provided as the **Heights** data set in the Analyst Sample Library. To open the **Heights** data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select **Heights**.
3. Click **OK** to create the sample data set in your **Sasuser** directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select **Sasuser** from the list of **Libraries**.
6. Select **Heights** from the list of members.
7. Click **OK** to bring the **Heights** data set into the data table.

### **Specify the Mixed Models Analysis**

To request a mixed models analysis, follow these steps:

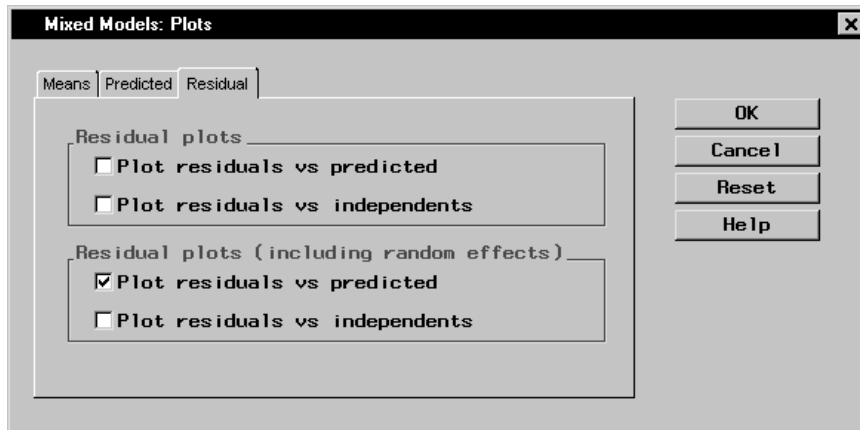
1. Select **Statistics** → **ANOVA** → **Mixed Models . . .**
2. Select **Height** as the dependent variable.
3. Select **Family** and **Gender** as classification variables.
4. Click **Model** to open the **Model** dialog.
5. Ensure that the **Fixed effects** check box is selected.
6. Select **Gender** and click **Add**.
7. Select the **Random effects** check box, and then select **Family** and click **Add**.
8. Select **Family** and **Gender**, and click **Cross**.
9. Click **OK** to return to the main dialog.

Based on your selections, the Mixed Models task constructs the **X** matrix by creating indicator variables for the **Gender** effect and including a column of 1s to model the global intercept. The **Z** matrix contains indicator variables for both the **Family** effect and the **Family\*Gender** interaction.

### **Produce a Residual Plot**

The Mixed Models task can produce means plots for fixed main effects and interactions, plots of predicted values, and residual plots that include or do not include random effects. To produce a plot of residuals versus predicted values that includes random effects, follow these steps:

1. Click **Plots** to open the **Plots** dialog.
2. Click on the **Residual** tab, and select **Plot residuals vs predicted** in the **Residual plots (including random effects)** box.

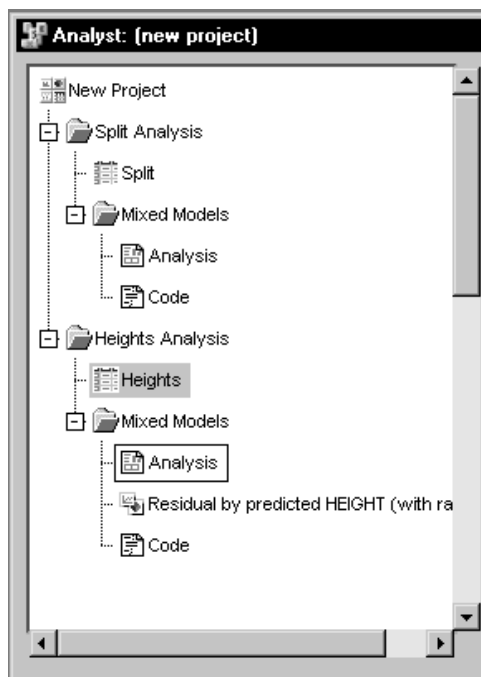


**Figure 15.9.** Mixed Model: Plots Dialog

When you have completed your selections, click **OK** in the main dialog to perform the analysis.

### **Review the Results**

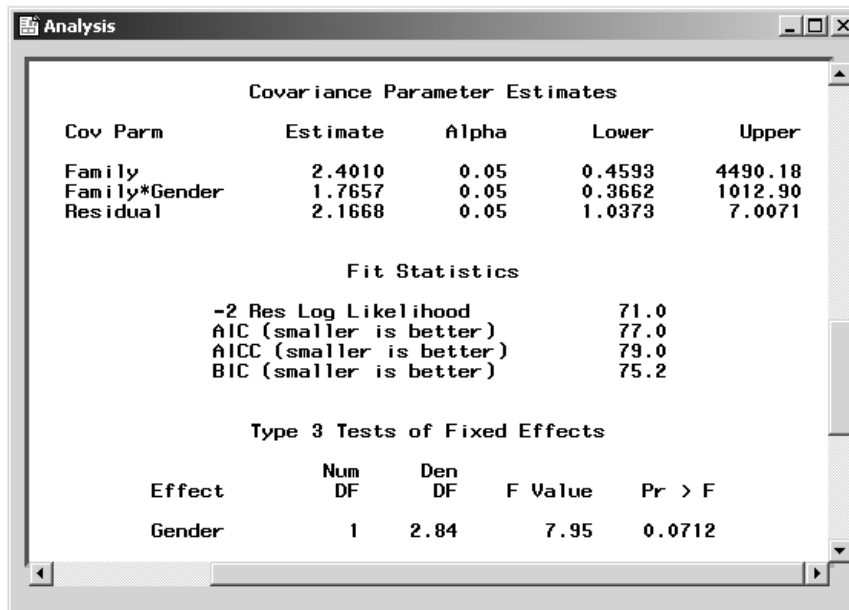
The results are presented in the project tree under the **Heights** data in the **Mixed Models** folder, as displayed in [Figure 15.10](#). The three nodes represent the mixed models results, the plot of residuals versus predicted values, and the SAS programming statements (labeled **Code**) that generate the output.



**Figure 15.10.** Mixed Models: Project Tree

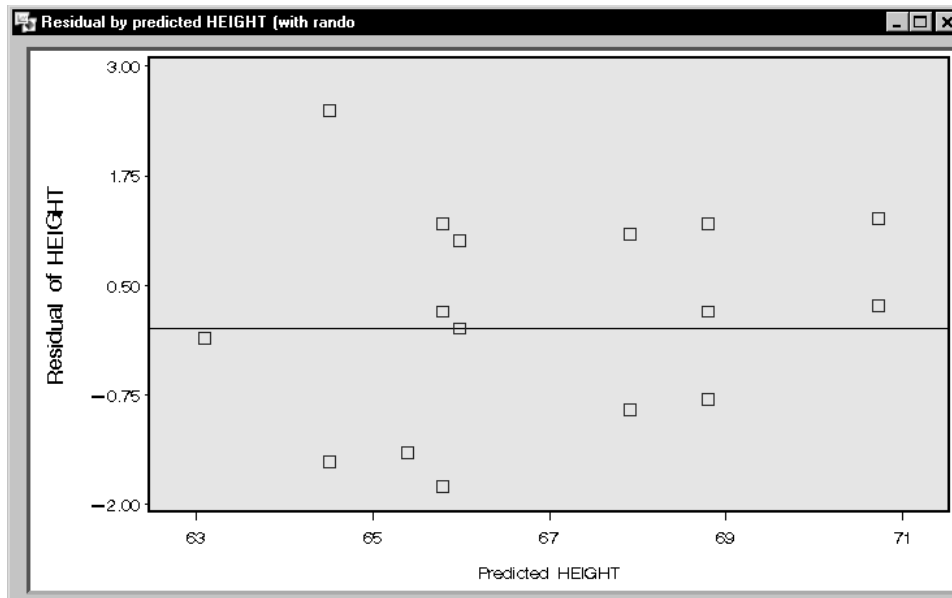
Double-click on the **Analysis** node in the project tree to view the contents in a separate window.





**Figure 15.11.** Mixed Models: Analysis Results

Figure 15.11 displays the mixed models analysis results for the clustered Heights data. The covariance parameter estimates for Family, Family\*Gender, and the residual variance are 2.4, 1.8, and 2.2, respectively. The “Test of Fixed Effects” table contains a significance test for the single fixed effect, Gender. With a  $p$ -value of 0.0712, the Type 3 test of Gender is not significant at the  $\alpha = 0.05$  level of significance. Note that the denominator degrees of freedom for the Type 3 test are computed using a general Satterthwaite approximation. A benefit of performing a random effects analysis using both Family and Family\*Gender as random effects is that you can make inferences about gender that apply to an entire population of families, not necessarily to the specific families in this study.



**Figure 15.12.** Mixed Models: Residuals Plot

Figure 15.12 displays a plot of the residuals versus predicted values that includes random effects,  $y - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\gamma}$  versus  $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}$ . Plots are useful for checking model assumptions and identifying potential outlying and influential observations. Based on the plot in Figure 15.12, the data seem to exhibit relatively constant variance across predicted values, and there do not appear to be any outliers or influential observations.

## References

- Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.
- Stroup, W. W. (1989), "Predictable Functions and Prediction Space in the Mixed Model Procedure," in *Applications of Mixed Models in*

*Agriculture and Related Disciplines*, Southern Cooperative Series  
Bulletin No. 343, Louisiana Agricultural Experiment Station, Baton  
Rouge, 39–48.



# Chapter 16

## Repeated Measures

### Chapter Contents

---

<b>Introduction</b> . . . . .	415
<b>Repeated Measures Analysis</b> . . . . .	416
<b>References</b> . . . . .	435



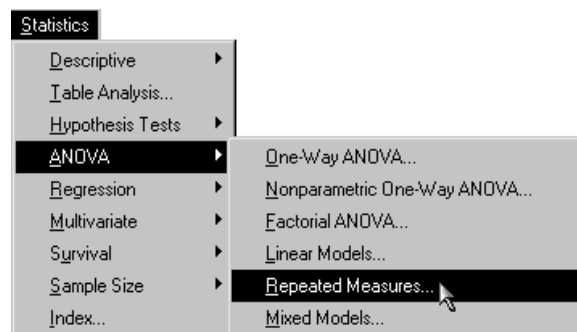
# Chapter 16

## Repeated Measures

---

### Introduction

Repeated measures analysis deals with response outcomes measured on the same experimental unit at different times or under different conditions. Longitudinal data are a common form of repeated measures in which measurements are recorded on individual subjects over a period of time. Blood pressure measured once a week for a month, CD4 counts tracked over a year in an AIDS clinical trial, and per capita demand deposits over years are examples of longitudinal data. Repeated measures can also refer to multiple measurements on an experimental unit, such as the thickness of vertebrae in animals.



**Figure 16.1.** Repeated Measures Menu

The experimental units are often subjects. In a repeated measurements analysis, you are usually interested in between-subject and within-subject effects. Between-subject effects are those whose values change only from subject to subject and remain the same for all observations on a single subject, for example, treatment and gender. Within-subject effects are those whose values may differ from measurement to measurement, for example, time. Usually, you are also interested in some between-subject and within-subject interaction, such as treatment by time.

Since measurements on the same experimental unit are likely to be correlated, repeated measurements analysis must account for that correlation. One way of doing this is by modeling the covariance structure of an individual's response. The *compound symmetry* structure assumes the same covariance between any two measurements and the same variance for each measurement. However, sometimes the covariance of measures that are close together in time is higher than the covariance for measurements further apart. In this case, the *first-order autoregressive* covariance structure may be more appropriate. Another possible covariance structure is *unstructured*, in which you estimate different parameters for the variance of each repeated measurement as well as different covariance parameters for each pair of repeated measurements.

The Repeated Measures task enables you to specify a repeated measures model with interactions and nested terms, define subject and repeated effects, and select from a wide range of covariance structures. You can estimate least-squares means for classification effects and output predicted values and residuals to a data set. Plots include means plots, predicted plots, and plots of residuals versus within and between effects. The Repeated Measures task applies methods based on the mixed model with special parametric structures on the covariance matrices.

The example in this chapter demonstrates how you can use the Repeated Measures task in the Analyst Application to analyze repeated measurements data.

---

## Repeated Measures Analysis

The data set analyzed in this task contains data from Littell, Freund, and Spector (1991). Subjects in the study participated in one of three different weightlifting programs, and their strength was measured once every other day for two weeks after they began the program. The first program increased the number of repetitions as the subject became stronger (RI), the second program increased the amount of weight as subjects became stronger (WI), and the subjects in the third program did not participate in weightlifting (CONT). The objective of this analysis is to investigate the effect each weightlifting program has on increasing strength over time. This section also illustrates how to prepare data in univariate form for this task.



### **Open the Weightsmult Data Set**

The data are provided in the Analyst Sample Library. To open the Weightsmult data set, follow these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select Weightsmult.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name . . .**
5. Select Sasuser from the list of **Libraries**.
6. Select Weightsmult from the list of members.
7. Click **OK** to bring the Weightsmult data set into the data table.

### **Data Management**

Figure 16.2 displays the Weightsmult data in multivariate form, which means that a single row in the data table contains all response measurements for a single subject. The **Program** variable defines the treatment group and takes the values 'CONT', 'RI', and 'WI'. The **Subject** variable defines each subject, and the variables **s1** through **s7** contain strength measurements across time for each subject.

	Subject	Program	s1	s2	s3	s4	s5	s6	s7
1	1	CONT	85	85	86	85	87	86	87
2	2	CONT	80	79	79	78	78	79	78
3	3	CONT	78	77	77	77	76	76	77
4	4	CONT	84	84	85	84	83	84	85
5	5	CONT	80	81	80	80	79	79	80
6	6	CONT	76	78	77	78	78	77	74
7	7	CONT	79	79	80	79	80	79	81
8	8	CONT	76	76	76	75	75	74	74
9	9	CONT	77	78	78	80	80	81	80
10	10	CONT	79	79	79	79	77	78	79
11	11	CONT	81	81	80	80	80	81	82
12	12	CONT	77	76	77	78	77	77	77
13	13	CONT	82	83	83	83	84	83	83
14	14	CONT	84	84	83	82	81	79	78
15	15	CONT	79	81	81	82	82	82	80
16	16	CONT	79	79	78	77	77	78	78

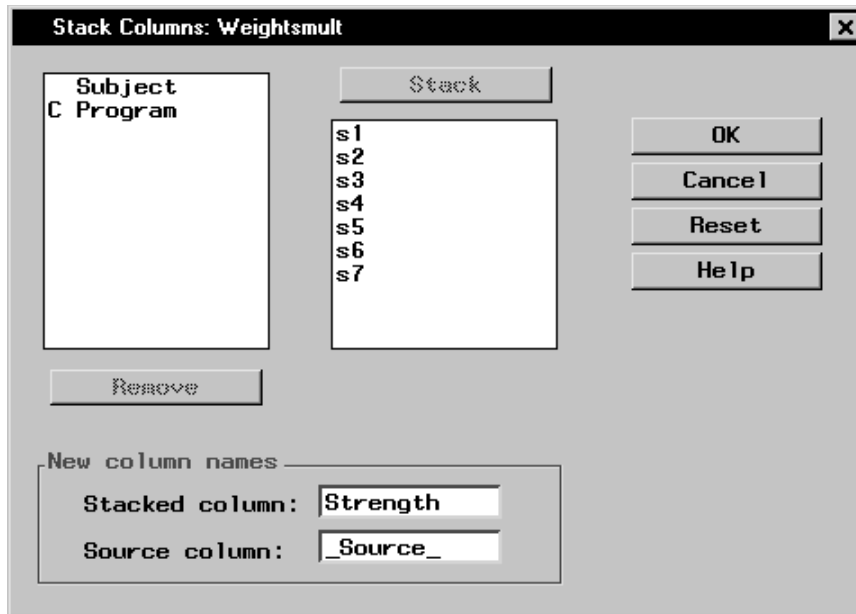
**Figure 16.2.** Weightsmult Data

In order for you to perform the repeated measures analysis using the Analyst Application, your data must be in univariate form, which means that each response measurement is contained in a separate row. If your data are not in univariate form, you must create a new data table with this structure. This can be accomplished via the Stack Columns task in the **Data** menu.

The Stack Columns task creates a new table by stacking specified columns into a single column. The values in the other columns are preserved in the new table, and a source column in the new data set contains the names of the columns in the original data set that contained the stacked values.

You want to put the values for columns corresponding to the strength measurement variables **s1** through **s7** in individual rows, so you want to stack columns **s1**–**s7**. To stack the columns, follow these steps:

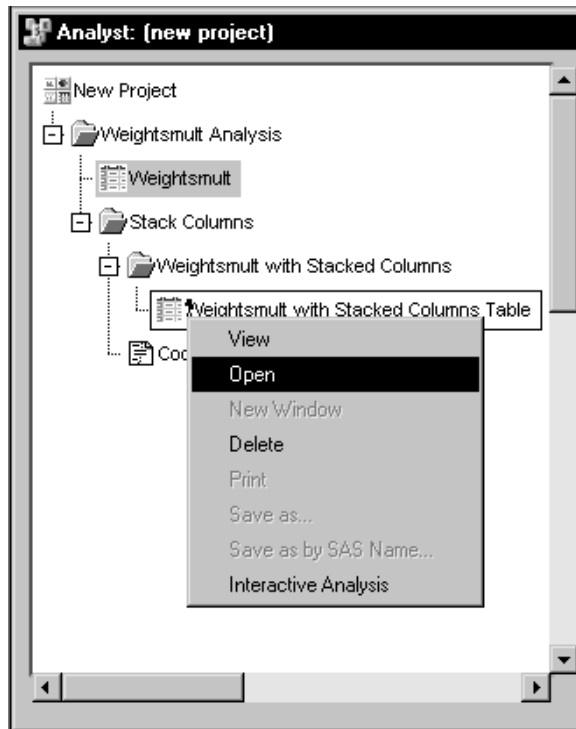
1. Select **Data** → **Stack Columns** . . .
2. Select **s1** through **s7** and click on the **Stack** button.
3. Type **Strength** in the **Stacked column:** field.
4. Click **OK** to produce the new data set.



**Figure 16.3.** Stack Columns Dialog

The new data set is presented in the project tree under the **Stack Columns** folder. The **Weightsmult with Stacked Columns** folder contains the new data set with the **Strength** stacked column, and the **Code** node contains the SAS programming statements that generated the data set.

If a view of the **Weightsmult with Stacked Columns** data is displayed, close it. Then right-click on the data set node labeled **Weightsmult with Stacked Columns**, as displayed in [Figure 16.4](#), and select **Open** to bring the new data set into the data table.



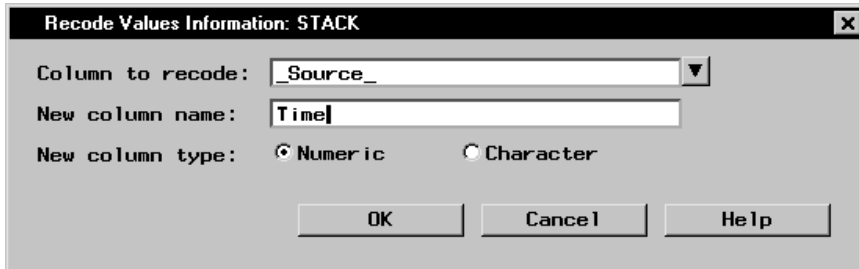
**Figure 16.4.** Stack Columns: Project Tree

The stacked columns data set contains two new variables. The **Strength** variable contains the strength measurements, and the **\_Source\_** variable denotes the measurement times with seven distinct character values: s1, s2, s3, s4, s5, s6, and s7. However, in this analysis, time needs to be numeric. You can create a numeric variable called **Time** by using the Recode Values facility.

To create the **Time** variable, follow these steps:

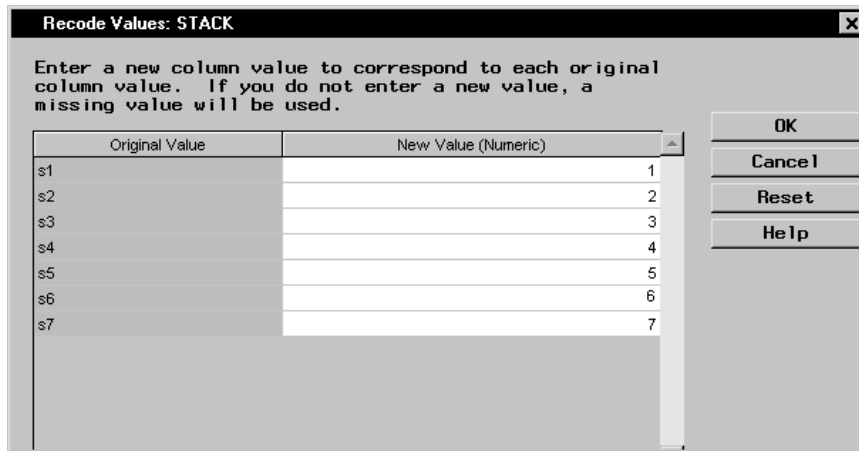
1. Select **Edit** → **Mode** → **Edit**.
2. Select **Data** → **Transform** → **Recode Values . . .**
3. Select **\_Source\_** as the **Column to recode**.
4. Type **Time** in the **New column name:** field.

5. Specify the new column type by selecting **Numeric**.
6. Click **OK** to enter values of the **Time** variable that correspond to current values of the **\_Source\_** variable.



**Figure 16.5.** Recode Values Information Dialog

7. Type **1** in the **New Value (Numeric)** column cell next to **s1**.
8. Type in the remaining numeric values corresponding to the original values of the **\_Source\_** column. [Figure 16.6](#) displays the final recoded values.
9. Click **OK** to create the new variable.



**Figure 16.6.** Recode Values Dialog

The data set now includes a variable **Time** that contains numeric values for the time of strength measurement. Because the time values are contained in a new variable, you can delete the original variable from the data set by right-clicking on the **\_Source\_** column in the data table and selecting **Delete**. Once you have deleted the column, the data set should contain four variables, **Subject**, **Program**, **Strength**, and **Time**, as displayed in [Figure 16.7](#).

	Subject	Program	Strength	Time
1	1	CONT	85	1
2	1	CONT	85	2
3	1	CONT	86	3
4	1	CONT	85	4
5	1	CONT	87	5
6	1	CONT	86	6
7	1	CONT	87	7
8	2	CONT	80	1
9	2	CONT	79	2
10	2	CONT	79	3
11	2	CONT	78	4
12	2	CONT	78	5
13	2	CONT	79	6
14	2	CONT	78	7
15	3	CONT	78	1
16	3	CONT	77	2
17	3	CONT	77	3
18	3	CONT	77	4
19	3	CONT	76	5

**Figure 16.7.** Weightsuni Data

Before proceeding with the analysis, you can save the new data set as **Weightsuni** by following these steps:

1. Select any cell in the data table or reselect the data set node labeled **Weightsmult** with **Stacked Columns** in the project tree.
2. Select **File** → **Save As By SAS Name . . .**
3. Type **Weightsuni** in the **Member Name** field and click **Save** to save the data set.

Note that the **Weightsuni** data are in univariate form and should be the same as the **Weights** data available in the Analyst Sample Library.

### Request the Repeated Measures Analysis

To specify the Repeated Measures task, follow these steps:

1. Select **Statistics** → **ANOVA** → **Repeated Measures . . .**
2. Select **Strength** as the dependent variable.
3. Select **Subject**, **Program**, and **Time** as classification variables.

Figure 16.8 displays the dialog with **Strength** specified as the dependent variable and **Subject**, **Program**, and **Time** specified as classification variables.

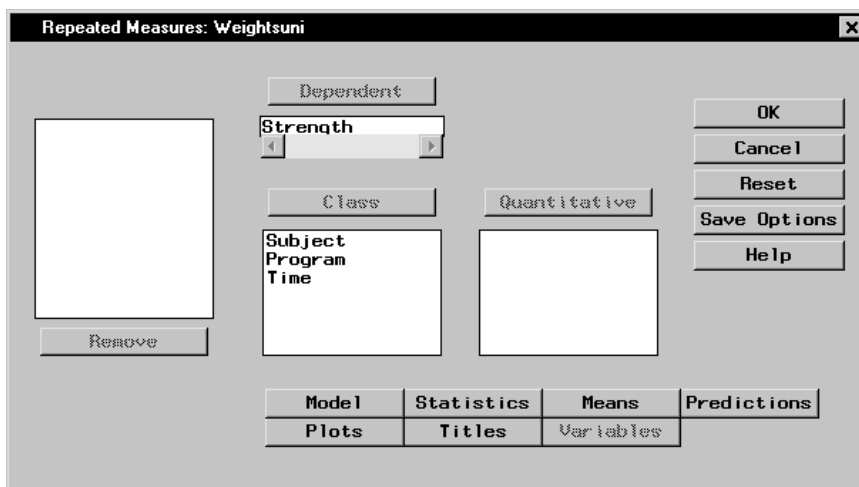


Figure 16.8. Repeated Measures Dialog

### Define the Model

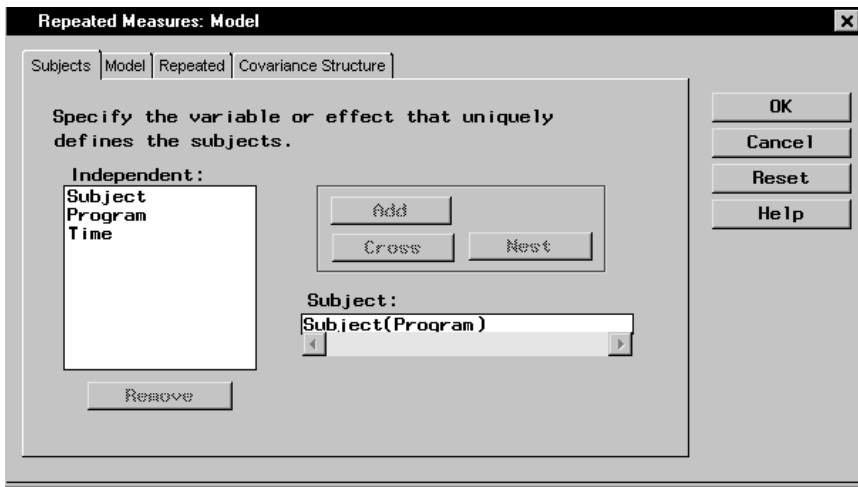
To perform a repeated measures analysis, you are required to specify a model, define subjects, specify a repeated effect, and select one or more structures for modeling the covariance of the repeated measurements. By defining a factorial structure between **Program** and **Time**, you can analyze the between-subject effect **Program**, the within-subject effect **Time**, and the interaction between **Program** and **Time**.



Each experimental unit, a subject, needs to be uniquely identified in the *Weightsuni* data set. The value of the **Subject** variable for the first subject in each separate **Program** is 1, the value of the **Subject** variable for the second subject in each **Program** is 2, and so on. Because subjects participating in different programs have the same value from the **Subject** variable, you need to nest **Subject** within **Program** to uniquely define each subject.

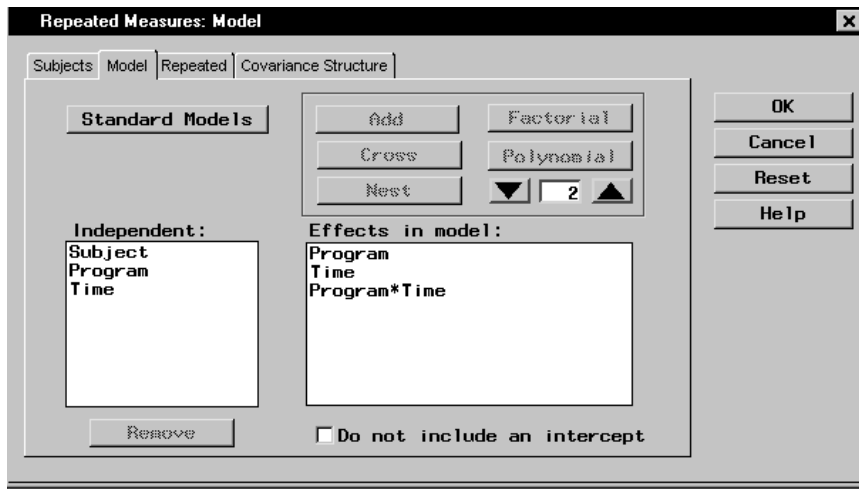
To define the repeated measures model, follow these steps:

1. Click on the **Model** button.
2. Select the **Subjects** tab.
3. Select **Subject** and click **Add**.
4. Select **Program** and click **Nest** to nest subjects within weightlifting programs.



**Figure 16.9.** Repeated Measures: Model Dialog, Subjects Tab

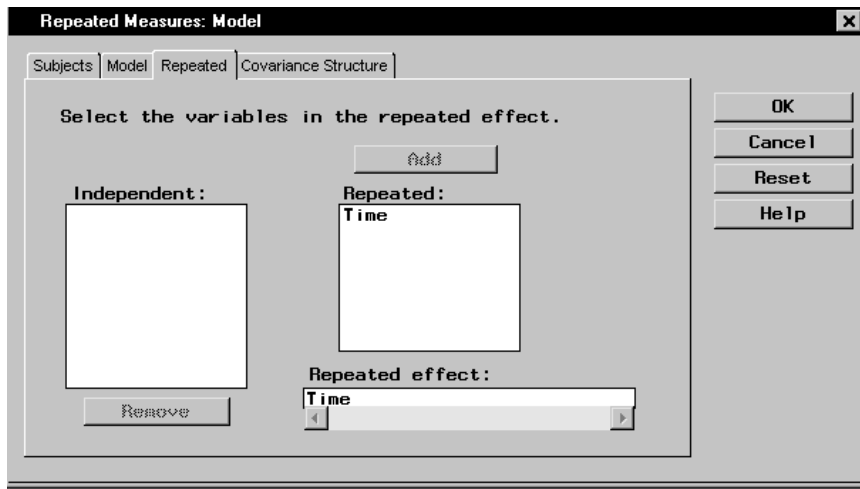
5. Select the **Model** tab.
6. Select **Program** and **Time** and click **Factorial** to specify a factorial arrangement, which is the main effects for **Program** and **Time** and their interaction.



**Figure 16.10.** Repeated Measures: Model Dialog, Model Tab

7. Select the **Repeated** tab.
8. Select **Time** and click **Add** to specify measurement times as the repeated effect.

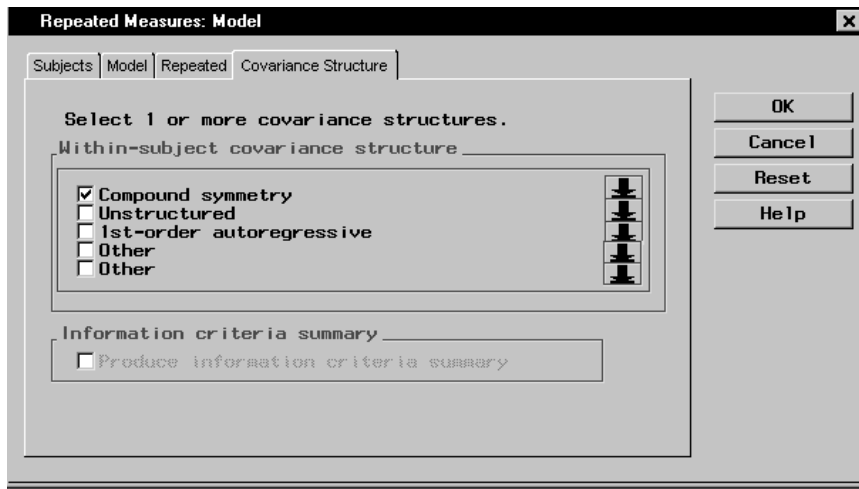
This identifies the repeated measurement effect.



**Figure 16.11.** Repeated Measures: Model Dialog, Repeated Tab

When analyzing repeated measures data, you must properly model the covariance structure within subjects to ensure that inferences about the mean are valid. Using the Repeated Measures task, you can select from a wide range of covariance types, where the most common types are compound symmetric, first-order autoregressive, and unstructured. To select the covariance structure for the analysis, follow these steps:

1. Select the **Covariance Structure** tab.
2. Select the **Compound symmetry** covariance structure.

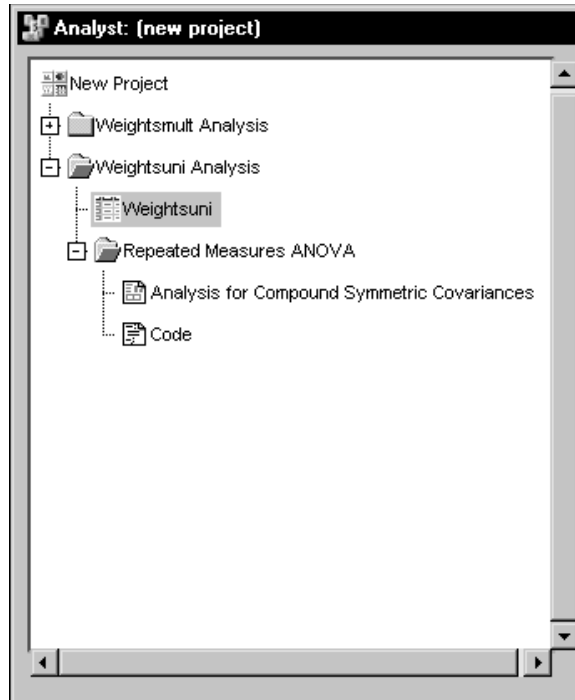


**Figure 16.12.** Repeated Measures: Model Dialog, Covariance Structure Tab

Close the Model dialog by clicking OK. When you have completed your selections, click **OK** in the main dialog to produce your analysis.

### **Review the Results**

The results are presented in the project tree under the **Repeated Measures ANOVA** folder, as displayed in [Figure 16.13](#). The nodes represent the repeated measures results and the SAS programming statements (labeled **Code**) that generated the output.



**Figure 16.13.** Repeated Measures: Project Tree

You can double-click on the **Analysis for Compound Symmetric Covariances** node in the project tree to view the results in a separate window.

The screenshot shows a window titled "Analysis for Compound Symmetric Covariances". Inside, there are two main sections: "Class Level Information" and "Dimensions".

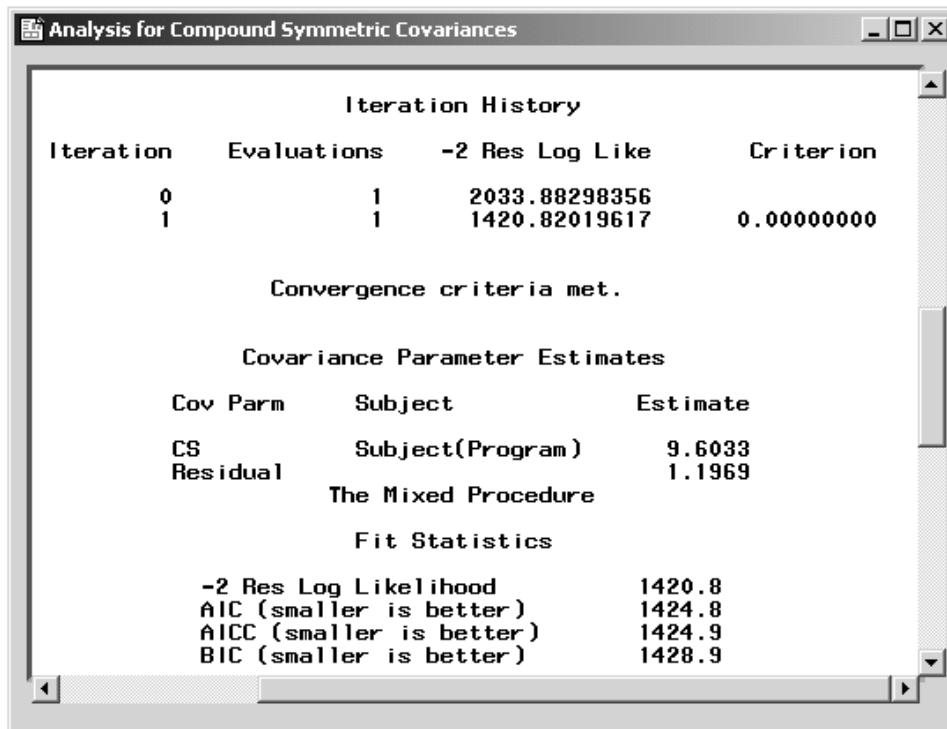
Class Level Information	
Class	Levels Values
Subject	21 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Program	3 CONT R1 W1
Time	7 1 2 3 4 5 6 7

Dimensions	
Covariance Parameters	2
Columns in X	32
Columns in Z	0
Subjects	57
Max Obs Per Subject	7
Observations Used	399
Observations Not Used	0
Total Observations	399

**Figure 16.14.** Repeated Measures: Model Information

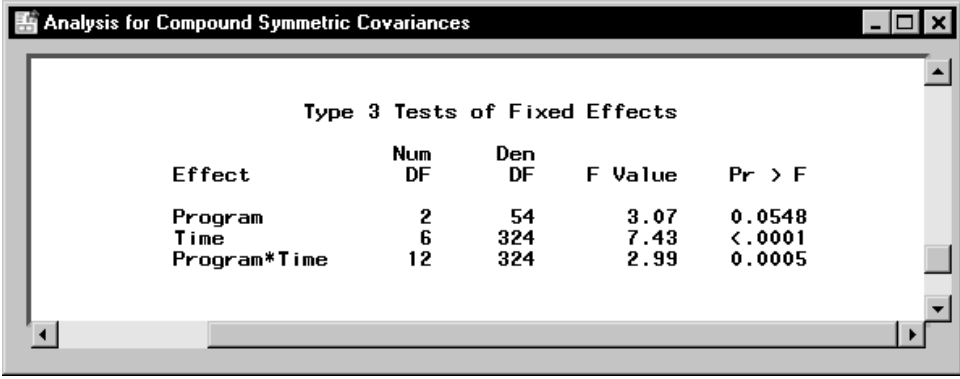
Figure 16.14 displays model information including the levels of each classification variable in the analysis. The **Program** variable has three levels while the **Time** variable has 7 levels. The “Dimensions” table displays information about the model and matrices used in the calculations. There are two covariance parameters estimated using the compound symmetry model: the variance of residual error and the covariance between two observations on the same subject. The 32 columns of the **X** matrix correspond to three columns for the **Program** variable, seven columns for the **Time** variable, 21 columns for the **Program\*Time** interaction, and a single column for the intercept. You should always review this information to ensure that the model has been specified correctly.



**Figure 16.15.** Repeated Measures: Fitting Information

Figure 16.15 displays fitting information, including the iteration history, covariance parameter estimates, and likelihood statistics. The “Iteration History” table shows the sequence of evaluations to obtain the restricted maximum likelihood estimates of the variance components.

The “Covariance Parameter Estimates” table displays estimates of the variance component parameters. The covariance between two measurements on the same subject is 9.6. Based on an estimated residual variance parameter of 1.2, the overall variance of a measurement is estimated to be  $9.6 + 1.2 = 10.8$ .



The screenshot shows a window titled "Analysis for Compound Symmetric Covariances". Inside the window, there is a table titled "Type 3 Tests of Fixed Effects". The table has five columns: "Effect", "Num DF", "Den DF", "F Value", and "Pr > F". The rows are "Program", "Time", and "Program\*Time".

Effect	Num DF	Den DF	F Value	Pr > F
Program	2	54	3.07	0.0548
Time	6	324	7.43	<.0001
Program*Time	12	324	2.99	0.0005

**Figure 16.16.** Repeated Measures: Tests for Fixed Effects

The “Type 3 Tests of Fixed Effects” table in [Figure 16.16](#) contains hypothesis tests for the significance of each of the fixed effects, that is, those effects you specify on the Model tab. Based on a  $p$ -value of 0.0005 for the **Program\*Time** interaction, there is significant evidence of a strong interaction between the weightlifting program and time of measurement at the  $\alpha = 0.05$  level of significance.

### Exploring Alternative Covariance Structures

Based on the assumption of the compound symmetry covariance structure, any two measurements on the same subject have the same covariance regardless of the length of the time interval between the measurements. However, repeated measurements are often more correlated when the measurements are closer in time than when they are farther apart. In this case, compound symmetry may not be appropriate, and you may want to investigate alternative covariance structures.

The first-order autoregressive covariance structure has the property that observations on the same subject that are closer in time are more highly correlated than measurements at times that are farther apart. The first-order autoregressive covariance can be represented by  $\sigma^2\rho^w$ , where  $w$  indicates the time between two measurements,  $\rho$  stands for the correlation between adjacent observations on the same subject, and  $\sigma^2$  stands for the variance of an



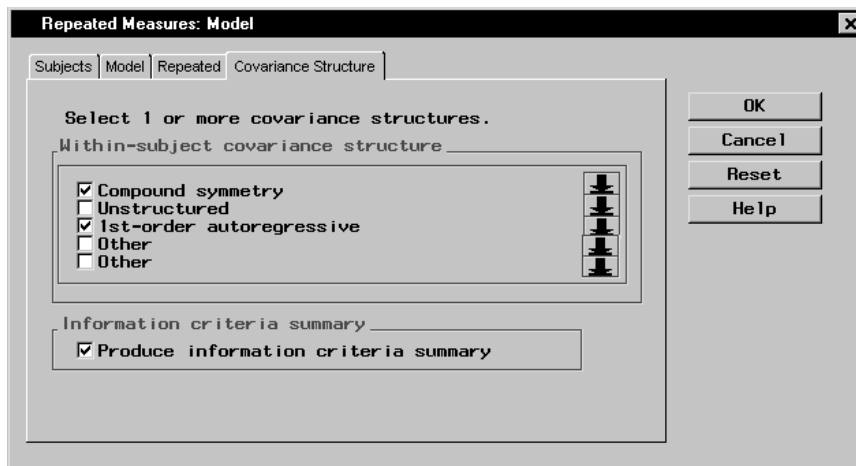
observation. For the first-order autoregressive covariance structure, the correlation between two measurements decreases exponentially as the length of time between the measurements increases.

To fit an additional repeated measures model with a first-order autoregressive covariance structure, follow these steps:

1. Select **Statistics** → **ANOVA** → **Repeated Measures . . .**

Note that the selections for the previous analysis are still specified.

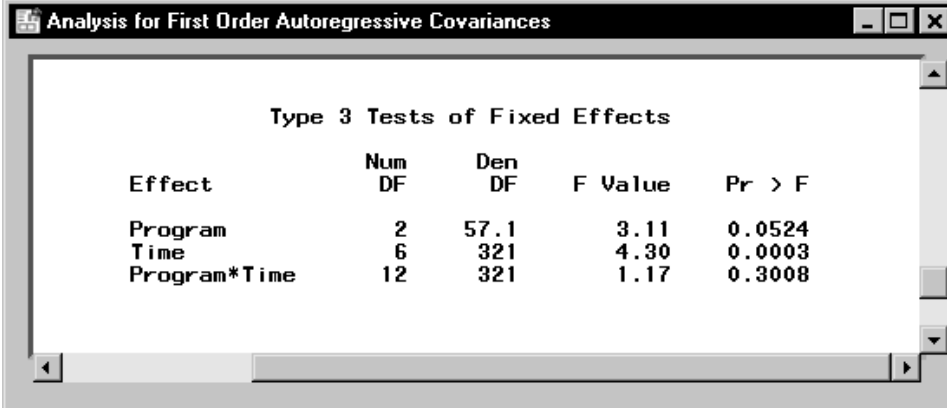
2. Click on the **Model** button.
3. Select the **Covariance Structure** tab.
4. Select the **1st-order autoregressive** structure.
5. Select **Provide information criteria summary** to produce a summary table of model-fit criteria for the two covariance structures.
6. Click **OK** in the main dialog to produce your analysis.



**Figure 16.17.** Repeated Measures: Model Dialog, Covariance Structure tab

Although this analysis models only two different covariance structures, the Analyst Application provides a wide range of structures to choose from, including unstructured, Huynh-Feldt, Toeplitz, and variance components. To select other structures, click on the down arrow next to an **Other** check box and choose from the resulting drop-down list.

Double-click on the **Analysis for First Order Autoregressive Covariances** node in the project tree to view the results in a separate window.



Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Program	2	57.1	3.11	0.0524
Time	6	321	4.30	0.0003
Program*Time	12	321	1.17	0.3008

**Figure 16.18.** Repeated Measures: Test for Fixed Effects for Autoregressive Covariance

Figure 16.18 displays the Type 3 tests for fixed effects based on the first-order autoregressive covariance model. Note that with a  $p$ -value greater than 0.30, the **Program\*Time** interaction is not significant at the  $\alpha = 0.05$  level of significance. The  $p$ -value is different from the  $p$ -value of the same test based on the compound symmetry covariance structure, and the two models lead to different conclusions. You can assess the model fit based on different covariance structures by comparing criteria that is provided in the Information Criteria Summary window in Figure 16.19.

Covariance Structure	Parameters	Akaike's Information Criterion	Schwarz's Bayesian Criterion	-2 Res Log Likelihood
Compound symmetry	2	1424.8	1428.9	1420.8
1st-order autoregressive	2	1270.8	1274.9	1266.8

**Figure 16.19.** Repeated Measures: Information Criteria Summary

The process of selecting the most appropriate covariance structure can be aided by comparing the Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (SBC) for each model. When you compare models with the same fixed effects but different variance structures, the models with the smallest AIC and SBC are deemed the best. In this example, the autoregressive model has lower values for both AIC and SBC, showing considerable improvement over the model with a compound symmetry structure. Based on the information criteria as well as the intuitively sensible property of the correlations being larger for nearby times than for far-apart times, the first-order autoregressive model is the more suitable fit for these data.

---

## References

- Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute Inc.
- Littell, R. C., Freund, R. J., and Spector, P. C. (1991), *SAS System for Linear Models, Third Edition*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (2000), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.



# Chapter 17

## Details

### Chapter Contents

---

<b>Customizing the Toolbar</b> . . . . .	439
Toolbars Tab . . . . .	439
Customize Tab . . . . .	441
<b>Resetting and Sharing Task Options</b> . . . . .	447



## Chapter 17

# Details

---

### Customizing the Toolbar

You can customize the Analyst toolbar to contain the tasks you use most often. You can also control icon size and toggle the display of tooltips and the toolbar.

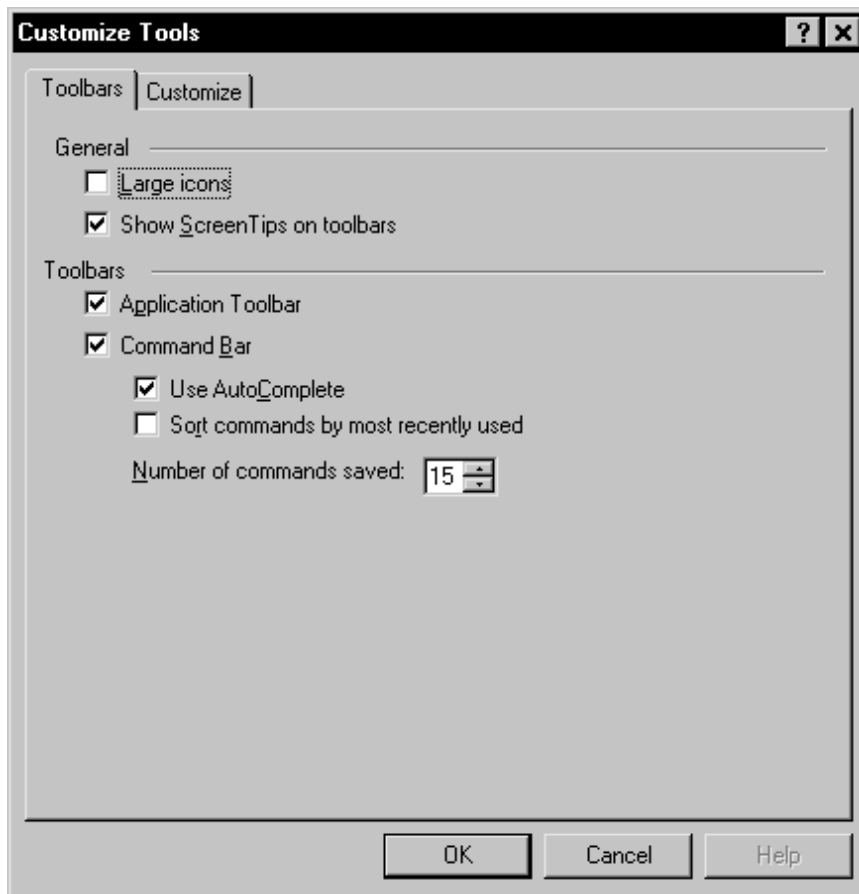
If you are on Windows, select **Tools** → **Customize...** to make changes to the Analyst toolbar. Under Unix, select **Options** from the **Tools** menu and select **Edit Toolbox** to display the Tool Editor dialog.

The following text refers to customizing the toolbar on the Windows operating system. Refer to the online help for specific information on customizing the toolbar on other operating systems.

---

### Toolbars Tab

In the **Toolbars** tab, you can specify general options that apply to the command bar and the toolbar.



**Figure 17.1.** Toolbars Tab

Under the **General** heading, click **Large icons** to display larger icons on the toolbar. If you leave **Large icons** unselected, the icons display in the default size.

Select **Show ScreenTips on toolbars** to display explanatory text when your cursor passes over an icon.



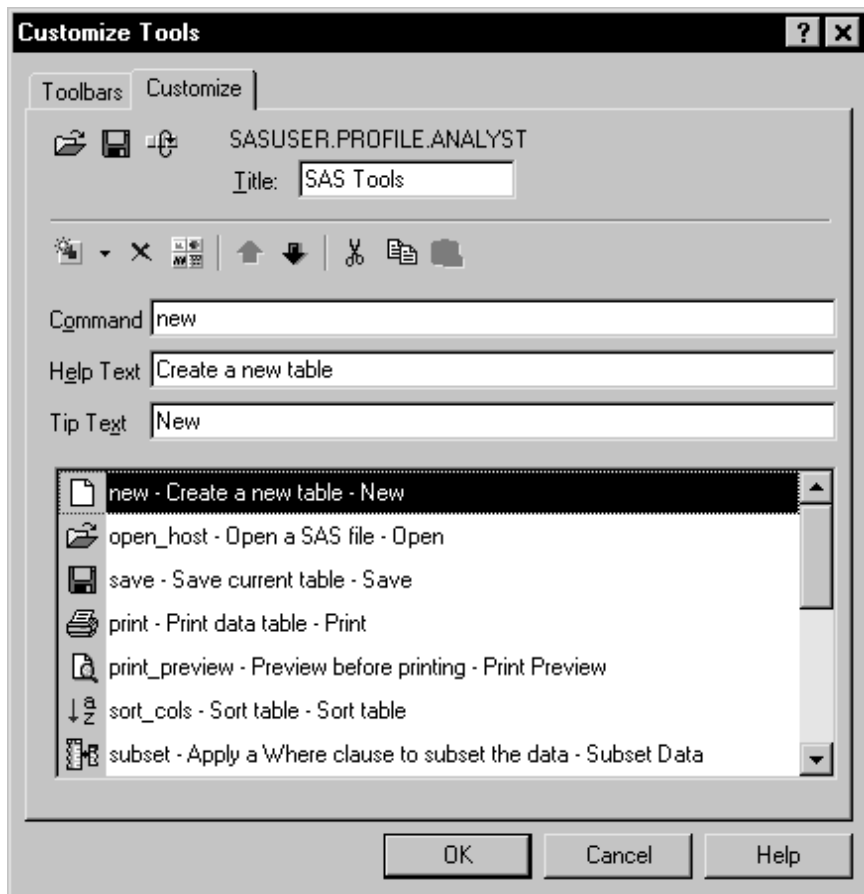
Under the **Toolbars** heading, select **Application Toolbar** to display the icons associated with any SAS window, including those of the Analyst Application. If **Application Toolbar** is unselected, no toolbar is displayed.

Select **Command Bar** to display the command bar from which you can issue SAS commands. If **Command Bar** is unselected, no command bar is displayed. Select **Sort commands by most recently used** to display the most recent commands at the top when you click on the arrow next to the command bar. Otherwise, commands are displayed in alphabetical order. Specify the **Number of commands saved** by clicking on the up or down arrow to change the number.

---

## Customize Tab

Click on the **Customize** tab to add or remove tasks from the toolbar, change the order of the icons, change the ScreenTip associated with an icon, or change the icon that is associated with a task.



**Figure 17.2.** Customize Tab

For more information about editing the Toolbar, click on the **Help** button in the Customize Tools dialog.

In order to add a task to the toolbar, you need to know the Analyst command for that task. The following tables list the command that is associated with each task.

**Table 17.1. File Commands**

<b>Task</b>	<b>Command</b>
New	NEW
Close	END
Open	OPEN_HOST
Open By SAS Name	OPEN_SAS
Open With New Query	QUERY_WINDOW
Open With Existing Query	QUERY_LIST
Save	SAVE
Save As	SAVEAS_HOST
Save As By SAS Name	SAVEAS_SAS
Projects	
New	NEW_PROJECT
Open	OPEN_PROJECT
Save	SAVE_PROJECT
Save As	SAVE_PROJECT_AS
Delete	DELETE_PROJECT
Print Preview	PRINT_PREVIEW
Print Setup	PRINT_SETUP
Print	PRINT

**Table 17.2. Edit Commands**

<b>Task</b>	<b>Command</b>
Insert Columns	ADDCOLS
Add Rows	ADDRWS
Duplicate	DUPLICATE
Delete	DELETE
Rename	RENAME
Mode	
Browse	BROWSE_MODE
Edit	EDIT_MEMBER_MODE
Shared Edit	EDIT_RECORD_MODE

**Table 17.3. View Commands**

<b>Task</b>	<b>Command</b>
Columns	
Move	MOVE_COLS
Hide	HIDE_COLS
Unhide	UNHIDE_COLS
Hold	HOLD_COLS
Labels	SHOW_LABELS
Table Attributes	TABATTRS

**Table 17.4. Tools Commands**

<b>Task</b>	<b>Command</b>
Titles	STITLES
Sample Data	SAMPLE_DATA
Viewer Settings	PREFS
Graph Settings	GRAPH_PREFS
New Library	LIB_ASSIGN

**Table 17.5. Data Commands**

<b>Task</b>	<b>Command</b>
Filter	
None	SUBSET_CLEAR
Subset Data	SUBSET
Sort	SORT_COLS
Transform	
Compute	COMPUTED_COLUMN
Rank	RANK
Standardize	STANDARDIZE
Recode Values	RECODE_VALUES
Recode Ranges	RECODE_RANGES
Convert Type	CONVERT_TYPE
Log(Y)	TRN_LOG
Sqrt(Y)	TRN_SQRT
1/Y	TRN_RECIP
Y*Y	TRN_SQUARE

**Table 17.5.** (continued)

<b>Task</b>	<b>Command</b>
Exp(Y)	TRN_EXP
Random Variates	
Normal	RV_NORMAL
Uniform	RV_UNI
Binomial	RV_BIN
Chi-Square	RV_CHI
Poisson	RV_POIS
Beta	RV_BETA
Exponential	RV_EXP
Gamma	RV_GAMMA
Geometric	RV_GEOM
Extreme Value	RV_EXTREME
Summarize By Group	SUM_BY_GROUP
Combine Tables	
Merge By Columns	MERGE
Concatenate By Rows	CONCATENATE
Stack Columns	STACK
Split Columns	SPLIT
Transpose	TRANSPOSE
Random Sample	RANDSAMP
Column Properties	COLATTRS

**Table 17.6. Reports Commands**

<b>Task</b>	<b>Command</b>
List Data	LIST_DATA
Tables	TABLES

**Table 17.7. Graphs Commands**

<b>Task</b>	<b>Command</b>
Bar Chart	
Horizontal	HBAR
Vertical	VBAR
Pie Chart	PIECHART

**Table 17.7.** (continued)

<b>Task</b>	<b>Command</b>
Histogram	HIST
Box Plot	BOX
Probability Plot	NORMPLOT
Scatter Plot	
Two-Dimensional	SCAT2D
Three-Dimensional	SCAT3D
Contour Plot	CONTOUR
Surface Plot	SURFACE

**Table 17.8.** Statistics Commands

<b>Task</b>	<b>Command</b>
Descriptive	
Summary Statistics	SUMMARY
Distributions	DISTRIB
Correlations	CORR
Frequency Counts	COUNTS
Table Analysis	TABLANAL
Hypothesis Tests	
One-Sample Z-test for a Mean	HT1Z
One-Sample t-test for a Mean	HT1T
One-Sample Test for a Proportion	HT1P
One-Sample Test for a Variance	HT1V
Two-Sample t-test for Means	HT2T
Two-Sample Paired t-test for Means	HT2PT
Two-Sample Test for Proportions	HT2P
Two-Sample Test for Variances	HT2V
ANOVA	
One-Way ANOVA	ONEANOVA
Nonparametric One-Way ANOVA	NONPARAM
Factorial ANOVA	FACANOVA
Linear Models	LINMOD
Repeated Measures	RMANOVA
Mixed Models	MIXED
Regression	

**Table 17.8.** (continued)

Task	Command
Simple	SIMPREGR
Linear	LINREGR
Logistic	LOGREGR
Multivariate	
Principal Components	PRINCOMP
Canonical Correlation	CANCORR
Survival	
Life Tables	LIFETEST
Proportional Hazards	PHREG
Sample Size	
One-Sample t-test	SSPMEAN1T
One-Sample Confidence Interval	SSPMEAN1CI
One-Sample Equivalence	SSPMEAN1E
Paired t-test	SSPMEANPT
Paired Confidence Interval	SSPMEANPCI
Paired Equivalence	SSPMEANPE
Two-Sample t-test	SSPMEAN2T
Two-Sample Confidence Interval	SSPMEAN2CI
Two-Sample Equivalence	SSPMEAN2E
One-Way ANOVA	SSPMEAN1A
Index	INDEX

**Table 17.9.** Help Command

Task	Command
Using This Window	window_help

---

## Resetting and Sharing Task Options

When you click on the **Save Options** button in a task dialog, the options that you set in that task are saved to an SLIST in the SASUSER.\_APPTSXS catalog. To restore all task settings to their defaults, remove the SASUSER.\_APPTSXS catalog. This removes any changes in options that you have made to all tasks. You can reset the options for a particular task

to their defaults by removing the SLIST from the SASUSER.\_APPTSKS catalog.

You can share your saved options by putting your SASUSER.\_APPTSKS catalog in a location where it can be copied. Other users who copy this catalog to their SASUSER directory have the same options set for all of their Analyst tasks.



# Appendix A

## Summary of Tasks

### Appendix Contents

---

<b>Reporting Tasks</b> . . . . .	451
<b>Graphical Tasks</b> . . . . .	452
<b>Statistical Tasks</b> . . . . .	456



## Appendix A

# Summary of Tasks

The following tables provide a list of capabilities available in the reporting, graphical, and statistical tasks in the Analyst Application. In each table, the Dialog column indicates the dialog in which the corresponding capability appears. Capabilities with an entry of *default* in the Dialog column are those that the task produces automatically.

Note that Analyst also provides an online index of its statistical features. You can view the index by clicking on the **Statistics** menu and selecting **Index**.

---

## Reporting Tasks

The following tables provide a list of capabilities available in the Analyst Application reporting tasks (**Reports** menu).

**Table A.1.** Capabilities in the List Data Task

Capability	Dialog
Column heading split character	Options
Column heading style	Options
Column values, row identifier	Main
Double spacing	Options
Sequence numbers, row identifier	Options
Single spacing	Options
Sum selected columns	Options
Total number of observations	Options

**Table A.2.** Capabilities in the Tables Tasks

Capability	Dialog
Cell format	Options
Formats for class values and statistics, supplied	Options
Formats for class values and statistics, user-defined	Options
Headings, empty class value combinations	Options

**Table A.2.** (continued)

<b>Capability</b>	<b>Dialog</b>
Labels, variables, and statistics	Options
Missing values as valid class levels	Options
Number of spaces, row titles	Options
Ordering, class values	Options
Summary column position	Options
Summary row position	Options
Text, empty cells	Options

---

## Graphical Tasks

The following tables provide a list of capabilities available in the Analyst Application graphical tasks (**Graphs** menu).

**Table A.3.** Capabilities in the Bar Chart Tasks

<b>Capability</b>	<b>Dialog</b>
Analysis variable	Options
Bar appearance	Options
Bar outline color and width	Options
Bar text color, size, and font	Options
Frame options	Options
Horizontal bar statistics, display options	Options
Number of bars	Options
Order of bars	Options
Reference lines	Options
Statistic to chart, average	Options
Statistic to chart, cumulative frequency	Options
Statistic to chart, cumulative percent	Options
Statistic to chart, frequency	Options
Statistic to chart, percent	Options
Statistic to chart, sum	Options
Three-dimensional chart	Main
Two-dimensional chart	Main
Vertical bar statistics, display options	Options

**Table A.4.** Capabilities in the Pie Chart Task

<b>Capability</b>	<b>Dialog</b>
Analysis variable	Options
Frequency variable	Options
Missing values	Options
Number of slices	Options
"Other" slice	Options
Slice and outline colors	Options
Slice angle	Options
Slice explosion	Options
Slice label type and placement	Options
Slice text color, size, and font	Options
Statistic to chart, average	Options
Statistic to chart, frequency	Options
Statistic to chart, percent	Options
Statistic to chart, sum	Options
Three-dimensional chart	Main
Two-dimensional chart	Main

**Table A.5.** Capabilities in the Histogram Task

<b>Capability</b>	<b>Dialog</b>
Bar and outline colors	Display
Bar pattern	Display
Exponential, fitted curve	Fit
Fitted curve colors	Display
Lognormal, fitted curve	Fit
Midpoints for histogram intervals	Display
Normal, fitted curve	Fit
Number of observations, vertical axis scale	Display
Percent of observations, vertical axis scale	Display
Proportion of observations, vertical axis scale	Display
Weibull, fitted curve	Fit

**Table A.6.** Capabilities in the Box Plot Task

Capability	Dialog
Box and outline colors	Display
Constant, box width	Display
Notches	Display
Point color and symbol	Display
Proportional to $\sqrt{n}$ , box width	Display
Proportional to $\log(n)$ , box width	Display
Proportional to sample size $n$ , box width	Display
Schematic style	Display
Skeletal style	Display

**Table A.7.** Capabilities in the Probability Plot Task

Capability	Dialog
Exponential, fitted curve	Main
Fitted curve color	Display
Fitted curve style and width	Display
Grid lines at percentiles	Display
Lognormal, fitted curve	Main
Normal, fitted curve	Main
Point color and symbol	Display
Weibull, fitted curve	Main

**Table A.8.** Capabilities in the Scatter Plot: Two-Dimensional Task

Capability	Dialog
Line color	Display
Line style and width	Display
Point color and symbol	Display
Points connected to $y = 0$	Display
Points connected with straight lines	Display
Reference lines	Display

**Table A.9.** Capabilities in the Scatter Plot: Three-Dimensional Task

Capability	Dialog
Point color and symbol	Display
Points connected to $x$ - $y$ plane	Display
Reference lines	Display
Rotation angle	Display
Tilt angle	Display

**Table A.10.** Capabilities in the Contour Plot Task

Capability	Dialog
Bivariate interpolation	Interpolate
Contour line labeling	Display
Interpolation / smoothing	Interpolate
Legend display	Display
Linear interpolation	Interpolate
Number of levels	Display
Partial spline interpolation	Interpolate
Pattern line density and angle	Display
Pattern outline color	Display
Pattern style	Display
Spline interpolation	Interpolate

**Table A.11.** Capabilities in the Surface Plot Task

Capability	Dialog
Bivariate interpolation	Interpolate
Interpolation / smoothing	Interpolate
Linear interpolation	Interpolate
Partial spline interpolation	Interpolate
Reference lines	Display
Rotation angle	Display
Spline interpolation	Interpolate
Surface colors	Display
Surface side walls	Display
Tilt angle	Display

## Statistical Tasks

The following tables provide a list of capabilities available in the Analyst Application statistical tasks (**Statistics** menu).

**Table A.12.** Capabilities in the Descriptive: Summary Statistics Task

Capability	Dialog
Box-and-whisker plot	Plots
Coefficient of variation	Statistics
Corrected sum of squares	Statistics
Histogram	Plots
Kurtosis	Statistics
Maximum	Statistics
Mean	Statistics
Median	Statistics
Minimum	Statistics
Number of missing observations	Statistics
Number of observations	Statistics
Output appearance	Output
Probability of $t$	Statistics
Range	Statistics
Skewness	Statistics
Standard deviation	Statistics
Standard error	Statistics
Student's $t$	Statistics
Sum	Statistics
Uncorrected sum of squares	Statistics
Variance	Statistics

**Table A.13.** Capabilities in the Descriptive: Distributions Task

Capability	Dialog
Box-and-whisker plot	Plots
Descriptive statistics	default
Exponential, fitted distribution	Fit
Extreme observations	default
Histogram	Plots



**Table A.13.** (continued)

<b>Capability</b>	<b>Dialog</b>
Lognormal, fitted distribution	Fit
Median	default
Moments	default
Normal, fitted distribution	Fit
Percentiles	default
Probability plot	Plots
Quantile-quantile plot	Plots
Quantiles	default
Sign statistic	default
Signed rank statistic	default
Tests for location	default
Weibull, fitted distribution	Fit

**Table A.14.** Capabilities in the Descriptive: Correlations Task

<b>Capability</b>	<b>Dialog</b>
Confidence ellipses	Plots
Corrected SSCP matrix	Options
Covariances	Options
Cronbach's alpha	Options
Descriptive statistics	Options
Hoeffding's D	Options
Kendall's tau- <i>b</i>	Options
<i>p</i> -values	Options
Pearson correlations	Options
Scatter plots	Plots
Spearman correlations	Options
SSCP matrix	Options

**Table A.15.** Capabilities in the Descriptive: Frequency Counts Task

<b>Capability</b>	<b>Dialog</b>
Bar charts	Plots
Cumulative frequencies	Tables
Cumulative percentages	Tables

**Table A.15.** (continued)

<b>Capability</b>	<b>Dialog</b>
Frequencies	Tables
Order, variable levels	Input
Percentages	Tables

**Table A.16.** Capabilities in the Table Analysis Task

<b>Capability</b>	<b>Dialog</b>
Chi-square statistics	Statistics
Fisher's exact test for $r \times c$ tables	Statistics
Frequencies	Tables
Likelihood ratio chi-square	Statistics
Mantel-Haenszel statistics	Statistics
McNemar's test for $2 \times 2$ tables	Statistics
Measures of agreement	Statistics
Measures of association	Statistics
Odds ratios for $2 \times 2$ tables	Statistics
Order, variable levels	Input
Pearson chi-square	Statistics
Pearson correlation coefficient	Statistics
Percentages	Tables
Simple kappa coefficient	Statistics
Spearman correlation coefficient	Statistics
Weighted kappa coefficient	Statistics

**Table A.17.** Capabilities in the Hypothesis Tests: One-Sample Z-test for a Mean Task

<b>Capability</b>	<b>Dialog</b>
Alternative hypotheses	Main
Bar chart	Plots
Box-and-whisker plot	Plots
Confidence intervals	Tests
Mean comparison value	Main
Normal distribution plot	Plots
Population standard deviation	Main

**Table A.17.** (continued)

<b>Capability</b>	<b>Dialog</b>
Population variance	Main
Power analysis	Tests

**Table A.18.** Capabilities in the Hypothesis Tests: One-Sample t-test for a Mean Task

<b>Capability</b>	<b>Dialog</b>
Alternative hypotheses	Main
Bar chart	Plots
Box-and-whisker plot	Plots
Confidence intervals	Tests
Mean comparison value	Main
Power analysis	Tests
$t$ distribution plot	Plots

**Table A.19.** Capabilities in the Hypothesis Tests: One-Sample Test for a Proportion Task

<b>Capability</b>	<b>Dialog</b>
Alternative hypotheses	Main
Bar chart	Plots
Confidence intervals	Tests
Normal distribution plot	Plots

**Table A.20.** Capabilities in the Hypothesis Tests: One-Sample Test for a Variance Task

<b>Capability</b>	<b>Dialog</b>
Alternative hypotheses	Main
Box-and-whisker plot	Plots
Confidence intervals	Tests
Probability distribution plot	Plots
Variance comparison value	Main

**Table A.21.** Capabilities in the Hypothesis Tests: Two-Sample t-test for Means Task

Capability	Dialog
Alternative hypotheses	Main
Bar chart	Plots
Box-and-whisker plot	Plots
Confidence intervals	Tests
Mean comparison value	Main
Means plot	Plots
Power analysis	Tests
Stacked data	Main
$t$ distribution plot	Plots
Unstacked data	Main

**Table A.22.** Capabilities in the Hypothesis Tests: Two-Sample Paired t-test for Means Task

Capability	Dialog
Alternative hypotheses	Main
Bar chart	Plots
Box-and-whisker plot	Plots
Confidence intervals	Tests
Mean comparison value	Main
Means plot	Plots
Power analysis	Tests
$t$ distribution plot	Plots

**Table A.23.** Capabilities in the Hypothesis Tests: Two-Sample Test for Proportions Task

Capability	Dialog
Alternative hypotheses	Main
Bar chart	Plots
Confidence intervals	Tests
Normal distribution plot	Plots
Stacked data	Main
Unstacked data	Main

**Table A.24.** Capabilities in the Hypothesis Tests: Two-Sample Test for Variances Task

Capability	Dialog
Alternative hypotheses	Main
Box-and-whisker plot	Plots
Confidence intervals	Tests
Probability distribution plot	Plots
Stacked data	Main
Unstacked data	Main

**Table A.25.** Capabilities in the ANOVA: One-Way ANOVA Task

Capability	Dialog
Bonferroni $t$ -test	Means
Box and whisker plot	Plots
Duncan multiple-range test	Means
Means comparisons	Means
Means plots	Plots
Power analysis	Tests
R-square statistic	default
Residual plots	Plots
Tests of homogeneity of variance	Tests
Tukey HSD test	Means
Welch's variance-weighted ANOVA	Tests

**Table A.26.** Capabilities in the ANOVA: Nonparametric One-Way ANOVA Task

Capability	Dialog
Ansari-Bradley test	Tests
Exact $p$ -values	Tests
Klotz test	Tests
Kruskal-Wallis test	Tests
Median test	Tests
Mood test	Tests
Savage test	Tests
Siegel-Tukey test	Tests

**Table A.26.** (continued)

<b>Capability</b>	<b>Dialog</b>
Van der Waerden test	Tests
Wilcoxon test	Tests

**Table A.27.** Capabilities in the ANOVA: Factorial ANOVA Task

<b>Capability</b>	<b>Dialog</b>
Adjusted R-square statistic	default
Bonferroni <i>t</i> -test	Means
Covariance ratio	Plots
Crossed effects	Model
DFFITS	Plots
Duncan multiple-range test	Means
Factorial models	Model
Influence plots	Plots
Interaction effects	Model
Least-squares means	Means
Leverage	Plots
Means comparisons	Means
Means plots	Plots
Model building	Model
Power analysis	Tests
Predicted values	Predictions
Prediction limits	Predictions
R-square statistic	default
Residual plots	Plots
Residual values	Predictions
Residuals, ordinary	Plots
Residuals, standardized	Plots
Residuals, studentized	Plots
Tukey HSD test	Means
Type 1, 2, 3, 4 sum of squares	Statistics
Weighted least squares	Tests

**Table A.28.** Capabilities in the ANOVA: Linear Models Task

<b>Capability</b>	<b>Dialog</b>
Adjusted R-square statistic	default
Bonferroni <i>t</i> -test	Means
Classification effects	Main
Covariance ratio	Plots
Crossed effects	Model
DFFITS	Plots
Duncan multiple-range test	Means
Factorial models	Model
Influence plots	Plots
Interaction effects	Model
Intercept	Model
Least-squares means	Means
Leverage	Plots
Means comparisons	Means
Means plots	Plots
Model building	Model
Multivariate tests	Tests
Nested effects	Model
Parameter estimates	Statistics
Polynomial effects	Model
Power analysis	Tests
Predicted plots	Plots
Predicted values	Predictions
Prediction limits	Predictions
R-square statistic	default
Residual plots	Plots
Residual values	Predictions
Residuals, ordinary	Plots
Residuals, standardized	Plots
Residuals, studentized	Plots
Scatter plots	Plots
Tukey HSD test	Means
Type 1, 2, 3, 4 sum of squares	Statistics
Weighted least squares	Tests

**Table A.29.** Capabilities in the ANOVA: Repeated Measures Task

<b>Capability</b>	<b>Dialog</b>
Ante-dependence covariances, first order	Model
Autoregressive covariances, first order	Model
Chi-square test, likelihood ratio	Statistics
Classification effects	Main
Compound symmetry covariances	Model
Confidence limits, covariance estimates	Statistics
Confidence limits, parameter estimates	Statistics
Covariance structures	Model
Crossed effects	Model
Factorial models	Model
Fitting information	default
Huynh-Feldt covariances	Model
Information criteria summary	Model
Interaction effects	Model
Intercept	Model
Least-squares means	Means
Likelihood ratio test	default
Means plots	Plots
Model building	Model
Nested effects	Model
Parameter estimates	Statistics
Polynomial effects	Model
Predicted plots	Plots
Predicted values	Predictions
Prediction limits	Predictions
Repeated effect	Model
Residual plots	Plots
Residual values	Predictions
Scatter plots	Plots
Subject effect	Model
Toeplitz covariances	Model
Type 1, 2, 3 sum of squares	Statistics
Unstructured covariances	Model
Variance components structure	Model



**Table A.30.** Capabilities in the ANOVA: Mixed Models Task

<b>Capability</b>	<b>Dialog</b>
Classification effects	Main
Confidence level	Options
Confidence limits, covariance parameter estimates	default
Confidence limits, fixed effects estimates	Options
Confidence limits, random effects estimates	Options
Covariance parameter estimates	default
Crossed effects	Model
Estimation methods	Options
Factorial models	Model
Fitting information	default
Fixed effects	Model
Interaction effects	Model
Intercept, fixed effects	Model
Least-squares means	Means
Main effects	Model
Maximum likelihood estimation	Options
Means plots, fixed effects	Plots
Minimum variance quadratic unbiased estimation	Options
Model building	Model
Nested effects	Model
Polynomial effects	Model
Predicted means	Predictions
Predicted value plots	Plots
Predicted values, including random effects	Predictions
Random effects	Model
REML	Options
Residual maximum likelihood estimation	Options
Residual plots	Plots
Satterthwaite method, fixed effects	default
Scatter plots	Plots
Solution, fixed effects parameters	Options
Solution, random effects parameters	Options
Types 1, 2, 3 estimation	Options
Types 1, 2, 3 tests, fixed effects	Tests
Variance components tests	Tests

**Table A.31.** Capabilities in the Regression: Simple Task

<b>Capability</b>	<b>Dialog</b>
Adjusted R-square statistic	default
Coefficient of variation	default
Confidence limits	Plots
Confidence limits for estimates	Statistics
Correlation matrix of estimates	Statistics
Covariance matrix of estimates	Statistics
Covariance ratio	Plots
Cubic model	Main
DFFITS	Plots
Influence plots	Plots
Leverage	Plots
Normal probability-probability plot	Plots
Normal quantile-quantile plot	Plots
Power analysis	Tests
Predicted values	Predictions
Prediction limits	Plots
Quadratic model	Main
R-square statistic	default
Residual plots	Plots
Residual values	Predictions
Residuals, ordinary	Plots
Residuals, standardized	Plots
Residuals, studentized	Plots
Scatter plots	Plots
Standardized regression coefficients	Statistics

**Table A.32.** Capabilities in the Regression: Linear Task

<b>Capability</b>	<b>Dialog</b>
Adjusted R-square model selection	Model
Adjusted R-square statistic	default
Akaike's information criterion	Model
Amemiya's prediction criterion	Model
Asymptotic covariance matrix	Statistics
Backward elimination model selection	Model

**Table A.32.** (continued)

<b>Capability</b>	<b>Dialog</b>
Bayesian information criterion	Model
Coefficient of variation	default
Collinearity analysis	Statistics
Confidence limits for estimates	Statistics
Correlation matrix of estimates	Statistics
Covariance matrix of estimates	Statistics
Covariance ratio	Plots
DFFITS	Plots
Durbin-Watson statistic	Statistics
Forward model selection	Model
Heteroscedasticity test	Statistics
Influence plots	Plots
Intercept	Model
Leverage	Plots
Mallows' Cp model selection	Model
Mallows' Cp statistic	Model
Maximum R-square improvement model selection	Model
Minimum R-square improvement model selection	Model
Multivariate statistics	Statistics
Normal probability-probability plot	Plots
Normal quantile-quantile plot	Plots
Partial correlations	Statistics
Power analysis	Tests
Predicted values	Predictions
Prediction limits	Plots
R-square model selection	Model
R-square statistic	default
Residual plots	Plots
Residual values	Predictions
Residuals, ordinary	Plots
Residuals, standardized	Plots
Residuals, studentized	Plots
Scatter plots	Plots
Schwarz's bayesian criterion	Model
Semi-partial correlations	Statistics

**Table A.32.** (continued)

<b>Capability</b>	<b>Dialog</b>
Standardized regression coefficients	Statistics
Stepwise model selection	Model
Stepwise regression	Model
Tolerance values for estimates	Statistics
Type 1 sum of squares	Statistics
Type 2 sum of squares	Statistics
Variance inflation factors	Statistics
Weighted least squares	Tests

**Table A.33.** Capabilities in the Regression: Logistic Task

<b>Capability</b>	<b>Dialog</b>
Association of predicted probabilities and observed responses	default
Backward elimination model selection	Model
Best subset model selection	Model
CI displacement	Plots
Classification effects	Main
Classification table	Statistics
Conditional odds ratios	Statistics
Confidence limits	Statistics
Correlation matrix of estimates	Statistics
Covariance matrix of estimates	Statistics
Crossed effects	Model
Deviance residuals	Plots
DFBetas	Plots
Difference in chi-square residuals	Plots
Difference in deviance residuals	Plots
Dispersion parameter	Statistics
Factorial models	Model
Fit statistics	default
Forward model selection	Model
Goodness-of-fit statistics	Statistics
Influence plots	Plots
Interaction effects	Model

**Table A.33.** (continued)

<b>Capability</b>	<b>Dialog</b>
Leverage	Plots
Likelihood ratio	default
Odds ratio estimates	default
Pearson residuals	Plots
Polynomial effects	Model
Predicted values	Predictions
Prior probabilities	Statistics
Probability cutpoints	Statistics
Profile likelihood limits	Statistics
Residual plots	Plots
Residual values	Predictions
Response profile	default
ROC curve	Plots
Standardized estimates	default
Stepwise model selection	Model
Wald limits	Statistics

**Table A.34.** Capabilities in the Multivariate: PrincipalComponents Task

<b>Capability</b>	<b>Dialog</b>
Analysis of correlation matrix	Statistics
Analysis of covariance matrix	Statistics
Analysis of uncorrected matrices	Statistics
Principal component scores	Save Data
Principal components plot	Plots
Scree plot	Plots

**Table A.35.** Capabilities in the Multivariate: CanonicalCorrelation Task

<b>Capability</b>	<b>Dialog</b>
Canonical redundancy statistics	Statistics
Canonical variable plot	Plots
Canonical variable scores	Save Data
Correlations of regression coefficients	Statistics
Number of canonical variables	Statistics

**Table A.35.** (continued)

<b>Capability</b>	<b>Dialog</b>
Partial correlations	Statistics
Partial variables	Variables
Regression analysis	Statistics
Semi-partial correlations	Statistics
Squared multiple correlation	Statistics
Standard error of coefficients	Statistics
Standardized regression coefficients	Statistics
<i>t</i> statistic and probability	Statistics

**Table A.36.** Capabilities in the Survival: Life Tables Task

<b>Capability</b>	<b>Dialog</b>
Censoring values	Main
Confidence intervals	Methods
Hazard function plots	Plots
Life table method	Methods
Probability density function plots	Plots
Product-limit estimation method	Methods
Strata endpoints	Plots
Survival estimates	default
Survival function plots	Plots

**Table A.37.** Capabilities in the Survival: Proportional Hazards Task

<b>Capability</b>	<b>Dialog</b>
Backward elimination model selection	Model
Best subset model selection	Model
Censoring values	Main
Confidence limits of hazard ratio	Methods
Correlations of parameter estimates	Methods
Covariances of parameter estimates	Methods
Failure time ties, Breslow approximate likelihood method	Methods
Failure time ties, discrete logistic model method	Methods
Failure time ties, Efron approximate likelihood method	Methods
Failure time ties, exact conditional probability method	Methods

**Table A.37.** (continued)

<b>Capability</b>	<b>Dialog</b>
Forward model selection	Model
Global hypothesis test	default
Stepwise model selection	Model
Survival function plots	Plots

The Sample Size tasks provide sample size and power calculations for several types of analyses and study designs. Power curves are available with each task. The types of sample size analyses available in the Analyst Application are as follows:

- one-sample  $t$ -test
- one-sample confidence interval
- one-sample equivalence
- paired  $t$ -test
- paired confidence interval
- paired equivalence
- two-sample  $t$ -test
- two-sample confidence interval
- two-sample equivalence
- one-way ANOVA





# Subject Index

## A

actuarial method,  
    See survival analysis, life-table  
    method  
adding rows, 41  
agreement, 261  
air quality data set, 270  
alphabetical order, 92  
alternative hypothesis, 211  
analysis of variance, 269  
    factorial ANOVA, 270, 284  
    linear models, 270, 290  
    mixed models, 270  
    multiple classification variables, 284  
    nonparametric one-way ANOVA, 269  
    one-way ANOVA, 269, 273  
    power analysis, 292, 295  
    quantitative variables, 290  
    repeated measures, 270  
    specifying interactions, 286, 291  
analysis results, printing, 78  
ANOVA, 22,  
    See analysis of variance  
association, 2×2 table, 238  
axis lines, 97

## B

Bandaid data set, 239  
bar charts  
    analysis variables, 113  
    bar values, 113  
    colors, 114  
    creating, 111  
    fonts, 114  
    frame options, 117

    frequency counts, 186, 189  
    number of bars, 112  
    reference lines, 117  
    statistics, 113, 115  
    three-dimensional, 112  
    two-dimensional, 112  
    variables, 118  
barcharts  
    titles, 117  
bars, options, 96  
between-subject effects, 415  
box-and-whisker plots  
    distributions, 196, 199  
    one-way ANOVA, 275, 279  
    paired t-test, 219, 222  
    summary statistics, 191, 194  
    two-sample test for variances, 229,  
        231  
browse mode, 33  
Bthdth92 data set, 184, 212, 218

## C

canonical coefficients, 372  
canonical correlations, 23, 357, 367, 373  
    eigenvalues, 368, 370  
    likelihood ratios, 368, 371  
    plots, 369, 375  
canonical variables, 372  
catalog entries, saving, 77  
censored observations, 382  
changing titles, 98  
charts  
    bar, 111  
    pie, 128  
classification variables, 44

- distributions, 199
  - summary statistics, 190
  - code
    - copying to Program Editor, 74
    - editing, 74
    - printing, 78
    - saving, 74
    - source, 93
    - submitting, 74
    - viewing, 73
  - color
    - in bar charts, 114
    - in pie charts, 131
    - in scatter plots, 142
  - columns
    - classes, 167
    - deleting, 39
    - displaying labels, 39
    - duplicating, 38
    - hiding, 35
    - holding, 36
    - inserting, 37
    - menu, 34
    - merging, 49
    - moving, 35
    - properties of, 40
    - sorting, 38
    - splitting, 51
  - combining tables, 48
  - commands, toolbar, 442
  - component loadings, 365
  - component plots, principal components, 360, 364
  - computing new variables, 43
  - concatenating rows, 48
  - confidence intervals
    - one-sample t-test, 214, 216
    - power calculations, 333, 348
    - sample size calculations, 333, 348
  - continuous variables, 44
  - Coronary2 data set, 317
  - correlations, 22, 184, 200
    - repeated measures analysis, 416, 432
    - scatter plots, 203, 206
    - types of, 184, 200
  - covariance estimates, mixed models, 403, 409
  - covariance matrix, principal components, 359
  - covariance types
    - first-order autoregressive, 432, 434
    - repeated measures analysis, 427, 432
  - Cox regression, 380, 390,
    - See also proportional hazards regression
  - creating projects, 69
  - Cronbach's alpha, 184
  - cumulative frequencies and percentages, 184, 188
  - customizing toolbar, 89, 439
- D**
- data
    - sorting, 42
  - data files, large, 91
  - Data menu, 17
    - computing log transformations, 47
    - computing new variables, 43
    - generating random variates, 47
    - recoding ranges, 44
  - data sets
    - air quality, 270
    - bandaid, 239
    - birth and infant mortality rates, 184, 212, 218
    - cereal judging, 53
    - coronary, 317
    - fitness, 200, 308
    - gym, 245
    - height, 405
    - houses, 302
    - infant mortality, 184, 212, 218
    - jobs, 367
    - opening, 28
    - protein, 358
    - rats survival, 381, 388
    - search algorithm, 224
    - split plot, 397
    - taste test, 53
    - test score, 227
    - weightlifting, 416

data table, 14  
     edit mode, 90  
     font, 90  
     modifying, 32  
     size of, 89  
 data views, opening, 28  
 data, editing, 33  
 deleting  
     columns, 39  
     project nodes, 72  
     projects, 73  
     rows, 42  
 descriptive statistics, 183  
     correlations, 184, 200  
     distributions, 184, 195  
     frequency counts, 184  
     summary statistics, 183, 190  
 distributions, 22, 184, 195  
     box-and-whisker plots, 196, 199  
     classification variables, 199  
     goodness of fit, 196, 198  
     histograms, 196, 199  
 duplicating  
     columns, 38  
     rows, 41  
**E**  
 Edit menu, 17  
 edit mode, 33, 90  
 editing code, 74  
 eigenvalues  
     canonical correlation, 368, 370  
     principal components, 359, 363  
 eigenvectors, principal components, 359, 363  
 equivalence tests, sample size and power, 338, 339  
     additive model, 338, 351  
     equivalence bounds, 340  
     multiplicative model, 339, 352  
     power calculations, 339  
 exact test, 245, 251  
 exponential, 43  
 Exposed data set, 381  
 expressions, 43

**F**  
 factorial ANOVA, 270, 284  
     means plots, 287, 289  
     specifying interactions, 286  
 fifth report style, 168  
 File menu, 16  
 file size, 91  
 files  
     local, 28  
     opening, 28  
 first report style, 162  
 Fitness data set, 200, 308  
 fixed effects,  
     See mixed models  
 folders  
     project, 13  
     renaming, 72  
 fonts, 90  
 format, listing report, 154  
 fourth report style, 167  
 frequency counts, 22, 184, 188  
**G**  
 goodness of fit, distributions, 196, 198  
 Gpa data set, 227  
 graph files, saving, 76  
 graph output, 93  
 Graph Settings, 94  
 graphs  
     axis lines, 97  
     bar options, 96  
     point display, 94  
     printing, 78  
     rectangle options, 96  
     text options, 97  
     types of, 20  
 Graphs menu, 17  
 Gym data set, 245  
**H**  
 Heights data set, 405  
 Help menu, 17  
 help, accessing, 19  
 hiding columns, 35  
 histograms, distributions, 196, 199  
 Hoeffding's D statistic, 184

holding columns, 36  
 homogeneity tests, survival analysis, 380, 386  
 Houses data set, 302  
 HTML output, 93  
 hypothesis tests, 22, 211  
   alternative hypothesis, 211  
   one-sample t-test, 212  
   paired t-test, 218  
   power calculations, 346  
   sample size calculations, 346  
   two-sample test for proportions, 224  
   two-sample test for variances, 227

**I**

icons, 439  
 index, 17  
 inserting columns, 37

**J**

Jobs data set, 367  
 JRATING data sets, 53

**K**

Kaplan-Meier method,  
   See survival analysis, product-limit method  
 Kendall's tau-*b*, 184

**L**

labels  
   displaying, 39  
   variables, 92  
 large data files, 91  
 least-squares means  
   defined, 400  
   mixed models, 404  
 life table method, 23, 380, 383, 385  
 likelihood ratios, canonical correlation, 368, 371  
 linear models, 270, 290  
   power analysis, 292, 295  
   scatter plots, 293, 297  
   specifying interactions, 291  
 linear regression, 23  
 lines, in scatter plots, 142

listing reports  
   creating, 153  
   format, 154  
   variables, 156  
 log transformations, 47  
 logarithm, 47  
 logistic regression, 23, 316,  
   See also regression  
   specifying interactions, 318  
 longitudinal data, repeated measures analysis, 415

**M**

means comparison test  
   one-way ANOVA, 274, 278  
   Tukey's studentized range test, 274, 278  
 means plots  
   factorial ANOVA, 287, 289  
   mixed models, 406, 410  
   paired t-test, 219, 223  
 menu  
   column, 34  
   Data, 17  
   Edit, 17  
   File, 16  
   Graphs, 17  
   Help, 17  
   Reports, 17  
   Statistics, 17  
   Tools, 17  
   View, 17  
 merging columns, 49  
 mixed models, 270, 395  
   clustered data, 405  
   compared to standard linear model, 395  
   covariance estimates, 403, 409  
   estimation methods, 396  
   fixed effects, 395  
   least-squares means, 400, 404  
   means plots, 406, 410  
   plots available, 396  
   random effects, 395  
   restricted maximum likelihood (REML) estimation, 396, 402

- specifying, 398, 406
- split plot design, 397
- moving columns, 35
- multiple linear regression, 307,
  - See also regression
  - collinearity analysis, 310, 314
  - confidence limits, 310, 313
  - residuals, 312
  - scatter plots, 311, 315
- multiple output, 93
- multivariate analysis, 357

**N**

- names of variables, 92
- nodes, 14, 16, 72
- nonparametric one-way ANOVA, 269

**O**

- one-sample t-test, 212
  - confidence intervals, 214, 216
  - t distribution plots, 215, 217
- one-sample test for a proportion, 232
- one-sample test for a variance, 232
- one-sample Z-test for a mean, 212, 231
- one-way ANOVA, 269, 273
  - box-and-whisker plots, 275, 279
  - means comparison test, 274, 278
  - power calculations, 343, 345, 348
  - sample size calculations, 343, 348
- opening
  - local files, 28
  - projects, 73
  - SAS data sets, 28, 29
  - SAS data views, 28, 29
- options
  - copying, 448
  - resetting, 448
  - saving, 98, 448
- output
  - graphs, 93
  - HTML, 93
  - multiple, 93

**P**

- paired samples, defined, 218

- paired t-test, 218
  - box-and-whisker plots, 219, 222
  - means plots, 219, 223
- partial correlation, principal components, 359
- Pearson correlations, 184
- pie charts
  - colors, 131
  - creating, 128
  - fonts, 131
  - frequency, 129
  - labels, 130
  - line width, 131
  - number of slices, 129
  - percent, 129
  - slice values, 129
  - three-dimensional, 128
  - titles, 134
  - two-dimensional, 128
  - variables, 134
- point display, 94
- power analysis, defined, 292
- power calculations, 327
  - confidence intervals, 333, 348
  - details, 346
  - equivalence tests, 339, 350
  - hypothesis tests, 329, 333, 346
  - one-way ANOVA, 343, 345, 348
  - precision, defined, 334
- power of a test, defined, 327
- principal components, 23, 357, 358
  - component plots, 360, 364
  - eigenvalues, eigenvectors, 359, 363
  - partial correlation, 359
  - scree plots, 359, 364
  - the covariance matrix, 359
- printing
  - code results, 78
  - graphs, 78
  - results, 78
- probability plots, 184
- project tree, size of, 89
- projects, 12
  - creating, 69
  - deleting, 73

- deleting nodes, 72
  - folder, 13
  - node, 14, 16
  - opening, 73
  - renaming, 71
  - saving, 70
  - saving under another name, 71
  - tree, 12
  - proportional hazards regression, 23, 380, 388, 390
    - model selection techniques, 380
  - Protein data set, 358
- Q**
- quantile-quantile plots, 184
  - Query window, 29
- R**
- random effects,
    - See mixed models
  - random sample, 43
  - random variates, 47
  - ranges, recoding, 44
  - ranking, 43
  - Rats data set, 388
  - reciprocal, 43
  - recode values, example, 420
  - recoding ranges, 44
  - recoding values, 43
  - rectangles, options, 96
  - regression, 301
    - collinearity analysis, 310, 314
    - confidence limits, 305, 310, 313
    - cubic model, 304
    - linear, 23
    - logistic, 23, 316
    - multiple linear, 307
    - quadratic model, 304
    - residuals, types of, 312
    - restricted maximum likelihood (REML) estimation, 396
    - scatter plots, 305, 306, 311, 315
    - simple linear, 302
  - renaming
    - folders, 72
    - projects, 71
  - repeated measures analysis, 270, 415
    - between-subject effects, 415
    - compound symmetry, 415, 430
    - covariance types, 427, 432
    - data format, 417
    - model specification, 424
    - plots available, 416
    - selecting appropriate covariance structure, 435
    - within-subject effects, 415
  - report styles
    - fifth, 168
    - first, 162
    - fourth, 167
    - second, 165
    - third, 166
  - reports
    - analysis variables, 163
    - column classes, 167
    - listing, 153
    - tabular, 162
    - titles, 156
    - types of, 20
  - Reports menu, 17
  - requirements, Analyst Application, iii
  - restricted maximum likelihood (REML) estimation, 396, 402
  - results
    - printing, 78
    - saving, 75
  - rotation angle, of scatter plots, 144
  - row classes, 166
  - rows
    - adding, 41
    - concatenating, 48
    - deleting, 42
    - duplicating, 41
- S**
- sample size, 23
  - sample size calculations, 327
    - confidence intervals, 333, 348
    - equivalence tests, 339, 350
    - hypothesis tests, 329, 333, 346
    - one-way ANOVA, 343, 348
  - SAS data sets, opening, 28, 29

- SAS data views, opening, 28, 29
- saving
  - catalog entries, 77
  - code, 74
  - graph files, 76
  - options, 98
  - projects, 70
  - results, 75
  - text files, 76
- scatter plots
  - connecting lines, 142
  - correlations, 203, 206
  - creating, 141
  - linear models, 293, 297
  - multiple linear regression, 311, 315
  - point appearance, 144
  - point color, 142
  - predicted versus observed, 293, 297
  - rotation angle, 144
  - simple linear regression, 305, 306
  - three-dimensional, 141
  - tick marks, 142, 144
  - tilt angle, 144
  - titles, 145
  - two-dimensional, 141
  - variables, 145
- scree plots, principal components, 359, 364
- Search data set, 224
- second report style, 165
- shared edit mode, 33
- simple linear regression, 302,
  - See also regression
  - confidence limits, 305
  - cubic model, 304
  - quadratic model, 304
  - scatter plots, 305, 306
- size, of data files, 91
- software requirements, iii
- sorting
  - columns, 38
  - data, 42
  - variables, 92
- source code, 93
- Spearman correlations, 184
- Split data set, 397
- splitting columns, 51
- SQL query, 29
- square, 43
- square root, 43
- stack columns, example, 418
- standardization, 43
- statistical tasks
  - ANOVA, 22, 269
  - canonical correlation, 23, 367
  - correlations, 22, 200
  - descriptive, 22, 183
  - distributions, 22, 195
  - frequency counts, 22, 184
  - hypothesis tests, 22, 211
  - life tables, 23, 380
  - linear regression, 23, 302, 303, 307
  - logistic regression, 23, 316
  - multivariate, 23, 357
  - power, 328
  - principal components, 23, 358
  - proportional hazards, 23, 380
  - repeated measures, 415
  - sample size, 23, 328
  - summary statistics, 22, 190
  - survival, 23, 379
  - table analysis, 22
  - types of, 21
- Statistics menu, 17
- styles
  - fifth report, 168
  - first report, 162
  - fourth report, 167
  - second report, 165
  - third report, 166
- submitting code, 74
- subsetting data, 52
- summarizing data, 43
- summary statistics, 22, 183, 190
  - box-and-whisker plots, 191, 194
  - classification variables, 190
- survival analysis, 23, 379
  - censored observations, 379, 382
  - component lifetimes, 379
  - hazards ratio, 390
  - life table method, 380, 383, 385

- Mantel-Haenszel test, 390
  - plots available, 380
  - product limit method, 379, 380
  - proportional hazards regression, 380, 388, 390
  - risk ratio, 390
  - SDF, survival distribution function, 379
  - survivor function plot, 383, 387
  - test for homogeneity, 380, 386
  - survival distribution function (SDF), 379
  - survivor function plot, 383, 387
- T**
- T distribution plots
    - one-sample t-test, 215, 217
  - table analysis, 22, 237
    - agreement, 263
    - association, 240, 247, 254
    - association in sets of tables, 251
    - chi-square test, 240, 247, 254
    - exact test, 249
    - Mantel-Haenszel statistics, 251
    - measures of association, 251
    - odds ratio, 251
  - tables, 48, 162
  - tabular reports
    - creating, 162
    - fifth report style, 168
    - first report style, 162
    - fourth report style, 167
    - second report style, 165
    - third report style, 166
  - tasks
    - accessing from index, 17
    - accessing from toolbar, 18
    - statistical, 21
  - text files, saving, 76
  - text, in graphs, 97
  - third report style, 166
  - three-dimensional bar charts, 112
  - tick marks, 142, 144
  - tilt angle, of scatter plots, 144
  - titles
    - changing, 98
  - toolbar, 18
    - commands, 442
    - customizing, 89, 439
  - toolbox, 18
  - Tools menu, 17
    - Graph Settings, 94
    - Viewer Settings, 89
  - tooltips, 439
  - transposing data, 43
  - Tukey's studentized range test, 274, 278
  - two-dimensional bar charts, 112
  - two-dimensional pie charts, 128
  - two-sample t-test for means, 232
  - two-sample test for proportions, 224
  - two-sample test for variances, 227
- V**
- variable labels, displaying, 92
  - variable names, displaying, 39, 92
  - variables
    - adding and removing, 16
    - classification, 44
    - computing, 43
    - continuous, 44
    - converting type, 43
    - display of, 92
    - sorting, 92
  - View menu, 17
  - Viewer Settings, 89
  - viewing code, 73
- W**
- Weights data set, 416
  - window layout, 89
  - within-subject effects, 415



# Syntax Index

## A

ANALYST, 1



# Your Turn

---

If you have comments or suggestions about *The Analyst Application, Second Edition*, please send them to us on a photocopy of this page or send us electronic mail.

For comments about this book, please return the photocopy to

SAS Publishing  
SAS Campus Drive  
Cary, NC 27513  
E-mail: [yourturn@sas.com](mailto:yourturn@sas.com)

For suggestions about the software, please return the photocopy to

SAS Institute Inc.  
Technical Support Division  
SAS Campus Drive  
Cary, NC 27513  
E-mail: [suggest@sas.com](mailto:suggest@sas.com)









